Polygraph

VOLUME 31 2002 NUMBER 2

Special Edition - Voice Stress
Contents

Published Quarterly

P.O. Box 8037, Chattanooga, Tennessee 37414-0037

Polygraph

Editor-in-Chief: Dean Pollina, Ph.D.

Managing Editor: Deedra Senter

******

Subscription Information: *Polygraph* is published quarterly by the American Polygraph Association. Advertising and Editorial Address is P.O. Box 10342, Ft. Jackson, SC 29207 (USA). Subscription Rates: One year $80.00 (domestic), $100.00 (foreign). Change of address: APA National Office, P.O. Box 8037, Chattanooga, TN 37414-0037. THE PUBLICATION OF AN ARTICLE IN POLYGRAPH DOES NOT CONSTITUTE AN OFFICIAL ENDORSEMENT BY THE AMERICAN POLYGRAPH ASSOCIATION.

The Sheridan Press

Acknowledgements

Paper presented at the 86th meeting of the Acoustical Society of America, Los Angeles, October 31, 1973. The research reported in this paper was conducted when the author was a graduate student in the Dept. of Psychology, University of Utah. Portions of this research were supported by a grant from the University of Utah Research Committee. Retyped in October 1977, incorporating slight changes. Address requests for reprints to: Gordon H. Barland, Ph.D., Barland & Associates, 565 East 4500 South, Salt Lake City, Utah 84107.

References

1. Barland, G.H. An experimental study of field techniques in lie detection. Unpublished Masters thesis, University of Utah, 1972, Available from the author. (a)

2. Barland, G.H. The reliability of polygraph chart evaluations. Polygraph 1972, 1, 192-206. To be published in N. Ansley (Ed.), Legal Admissibility of the Polygraph, Springfield, Ill.: C. C. Thomas, in press. (b)

3. Barland, G.H. & Raskin, D.C. Detection of deception. In W. F. Prokasy and D.C. Raskin (Eds), *Electrodermal activity in psychological research.* New York: Academic Press, 1973, 417-477.

4. Gustafson, L.A. & Orne, M.T. Effects of heightened motivation on the detection of deception. *Journal of applied psychology,* 1963, 47, 408-411.

5. Kubis, J.F. Department of Psychology, Fordham University, Bronx, New York. Personal communication, 1973.

6. Kugelmass, S. & Lieblich, I. The effects of realistic stress and procedural interference in experimental lie detection. *Journal of applied psychology,* 1966, 50, 211-216.

7. Lykken, D.T. The validity of the guilty knowledge technique: The effects of faking. *Journal of applied psychology,* I960, 44, 253-262.

8. Orne, M.T., Thackray, R.I., & Paskewitz, D.A. On the detection of deception: A model for the study of the physiological effects of psychological stimuli. In N. Greenfield & R. Sternbach (Eds)., *Handbook of psychophysiology* . New York: Holt, Rinehart & Winston, 1972, 743-785.

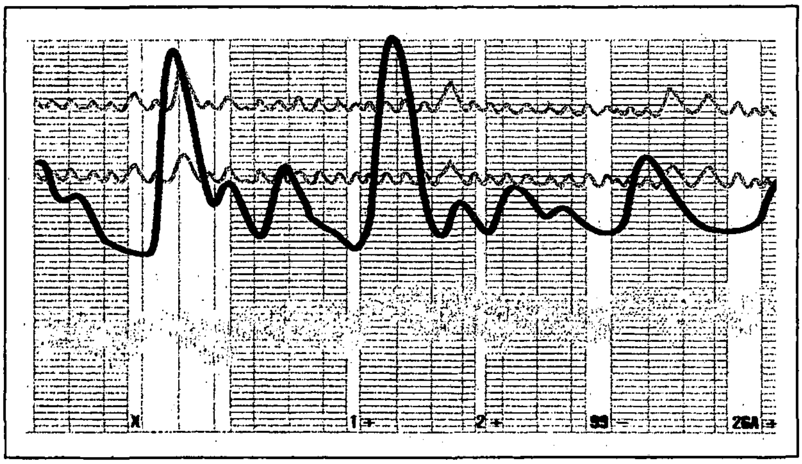
9. O'Toole, G. Assassination tapes. *Penthouse,* 1973, 4(11) 44-47, 112-114, 124, 126.

10. Simonov, P.V. & Frolov, M.V. Utilization of human voice for estimation of man's emotional stress and state of attention. *Aerospace Medicine.* 1973, 44, 256-258

11. Violante, R. & Ross, S.A. Research in interrogation procedures. Office of Naval Research, Report 707-65, filed with Defense Documentation Center, AD 467 624, October, 1964

12. Worth, J.W. Department of Psychology, Washington and Lee University, Lexington, Virginia. Personal communication, August, 1973.

Academy of Polygraph Science



The Academy of Polygraph Science conducts certification training in basic and advanced forensic psychophysiology in the detection of deception courses.

The academy was founded to provide quality polygraph training to qualified highly motivated individuals in law enforcement, government and the private sectors.

Academy accreditations include the American Polygraph Association(APA),Florida Polygraph Association(FPA) and the American Association of Police Polygraphist(AAPP). We comply with all ASTM standards.

We also conduct advanced training in Post Conviction Sex Offender Testing.

Contact us for more information on upcoming basic and advanced classes.

Dr. Richard E. Poe
Director

Dr. Richard E. Poe has been studying and practicing Forensic Psychophysiology for more than 20 years. He is a graduate of the University of Sarasota, in Human Services. Prior to that, he completed his Masters degree and Batchelor's degree's in business management. He is seasoned law enforcement officer and polygraphist .He has established a private practice in Largo,Fl.where he continues to practice and teach. Dr. Poe is the current Vice President of the Florida Polygraph Assoc . and holds F.P.A. certified polygraph certificate #127.He also holds an American Assoc. of Police Polygraphist certificate #1745.As well as being certified by APA and FPA in Post Conviction Sexual Offender Testing.

PHONE: (727) 584-5388 E MAIL : ACDYPOLYSCIENCE@IX.NETCOM

FAX: (727) 584-4339 WEBSITE: WWW.DRPOEANDASSOC.COM
MAILING ADDRESS: 12945 Seminole Blvd. Bldg #1 Suite #4 Largo, Fl . 33778-2313

TABLE 6
Individual Components Ranked for Effectiveness on each Criminal Suspect

Subject	Respiration	SRR	Cardiovascular	Voice
1	4[a]	1	3	2
2	4	3	1	2
3	4	2	3	1
4	3	1	(4)b	2
5	1	2	(4)	3
6	4	2	3	1
7	4	1	2	3
8	2	1	(4)	3
9	1	2	3.5	3.5
10	(3.5)	1	2	(3.5)
11	1	3	(4)	2
12	2	1	3.5	3.5
13	2.5	1	2.5	4
14	2	1	4	3
Mean Ranks	2.71	1.57	2.89	2.46

[a] Rank of 1 = most effective component, 4 = least effective component

[b] Bracketed ranks indicate disagreement between that component and the total polygraph score.

knowledge of the polygraph outcome influenced the interpretation of the PSE charts. Previous research at our laboratory indicates that this is not a significant source of bias when the charts are being evaluated numerically (Barland, 1972b). However, to serve as a check on this possibility the PSE charts were interpreted completely in the blind by one of the inventors of the PSE. The blind evaluator did not know what questions had been asked or what each case involved. He was merely informed as to which were the relevant questions and which were the centrol questions. Since he was not familiar with the numerical evaluation technique, he made dichotomous decisions of "deceptive" vs "not deceptive." He was instructed that he could make a third choice of "inconclusive", but he preferred to make a definite decision in every case.

The blind evaluator made 11 decisions of "deceptive" and 3 decisions of "not deceptive" (i.e., 3 disagreements with the polygraph). Use of the binomial model (14, ½) found this to be significant at the .05 level. In two of the three cases where the blind PSE

TABLE 7
Numbers of Inconclusive and Errors for each Physiological Parameter at varying cut-off points for the Inconclusive Region

Inclusive Boundaries of the inconclusive region	Respiration		SRR		Cardiovascular		Voice	
	No. Incl.	No. Errors	No. Incl.	No. Errors	No. Incl.	No. Errors	No. Incl.	No. Error
0	5	1 *[a]	0	0 ***	3	3 n.s.	1	1 **
± 1	7	0 **	0	0 ***	7	2 n.s.	4	0 ***
± 2	7	0 **	1	) ***	9	2 n.s.	6	0 **

*	P< .05
**	P< .01
***	P< .001
n.s.	not significant

[a] Level of significance was determined by a binomial test on the number of errors out of the number of decisions.

evaluation disagreed with the polygraph, my own evaluation of the PSE charts had resulted in a decision of inconclusive. Thus in the 8 cases where both of us had made decisions based upon the voice alone, there was only one disagreement. This was significant at the .05 level using the binomial model (8, ½).

It should be noted that in this high-stress study, the voice data were obtained simultaneously with the polygraph data. Thus, the test was structured around the polygraph technique. The suspects' replies were either "yes" or "no" rather than explanatory, and there were pauses of about 20 seconds between each of the replies. Moreover, the Subjects experienced some degree of discomfort from the blood pressure cuff as the polygraph examination proceeded. This could be expected to induce a certain amount of stress into the testing situation which, by increasing the base level of stress, would tend to mask the responses caused by lying. It would be reasonable to hypothesize that the efficiency of voice analysis in a lie detection situation would be higher in a situation structured around the voice technique.

The results of the high-stress study shows that reliable changes occur in the voice which are correlates of short-term psychological stress evidenced by changes in the autonomic nervous system. Yet no significant results were obtained in the low stress study. This suggests the hypothesis that a certain amount of stress must be reached within an individual before reliable stress-related changes occur in the voice.

The difference in the level of stress between these two studies reported here is not the only difference between the two experiments. The two Subject populations were very different in a number of ways: age, education, socio-economic status, number of arrests, IQ, etc. Another difference was the testing methodology used. In the low stress experiment a peak of tension was used, whereas in the high stress experiment a control question test was used. A final difference was that the subjects in the low-stress experiment were instructed to try hard to beat the test and to keep their voice the same each time in order not to give it away in their voice. Previous research has indicated that the harder a Subject tries to beat the test, the easier his lies are detected (Gustafson and Orne, 1963, Lykken, I960). Perhaps the voice, unlike autonomic indices, is more amenable to voluntary control. However, recent unpublished work by Worth (personal communication, 1973) supports the view that the level of stress experienced by the subject is an important factor affecting the accuracy of voice analysis in a lie detection situation. This hypothesis will be investigated further.

activity by means of occlusion plethysmography. The verbal answers to the test questions were simultaneously recorded for later analysis by the same equipment used in Experiment 1.

All suspects were administered the federal government's modification of the zone comparison polygraph test. This test consists of a series of at least 10 questions of which three are relevant questions pertaining to the incident under investigation, and three are control questions, designed to cause the *innocent person to respond. The response to each* relevant question is compared to the response to its adjacent control question, and the pair of responses is given a numerical score ranging from a +3 to a -3. If the two responses are of about equal magnitude or are nonexistent, the pair of questions is scored 0. A plus indicates that the person responded more to the control question than to the relevant: a minus indicates the opposite. The value of 1, 2, or 3 indicates the degree of inequality between the two responses. This evaluation is made for each component (respiration, Skin Resistance Response, and cardiovascular activity) on each pair of questions. The list of ten questions is repeated for a minimum of at least three trials. The zone comparison test takes about three hours to administer and is described in detail elsewhere (Barland & Raskin, 1973).

All of the scores on the polygraph from one suspect were summed. If the total score was +6 or higher, it was concluded that the suspect told the truth; if it was between + or - 5, inclusive, the result was inconclusive; if it was -6 or lower, it was concluded that the suspect lied on the test. The numerical scoring system has been found to be both valid and reliable, and has been described more fully elsewhere (Barland, 1972a, 1972b).

The polygraph examiner concluded that all 14 suspects had lied when they denied involvement in the crime. No inconclusive polygraph examinations were included in the sample. There is thus good reason to believe that all of the suspects had been under a high degree of stress when they answered the relevant questions. Although the examiner's decision was completely confirmed in 6 of the cases and there is no reason to believe that any of the decision were wrong, the issue of whether the polygraph examiner's decisions were all correct is not important. The important thing is that the suspects had shown a stronger physiological arousal, as measured by the polygraph, when they answered the relevant questions than they did when they answered the control questions. Since it has long been established that the polygraph is highly effective in measuring short-term psychological stress in lie detection situations, the question explored in this study is the extent to which autonomic changes recorded by the polygraph will be reflected by changes in the voice.

Approximately one week after each polygraph examination, the tape recording was analyzed on the PSE-1. Two analyses were made: Mode 1 at 7½ ips and Mode III at either 1-7/8 or 15/16th ips, depending upon the type of pattern obtained. The two sets of PSE charts were then numerically scored in the same manner as the polygraph charts had been: each pair of control and relevant questions was rated on a 7-point scale ranging from +3 to -3, and all of the individual scores thus obtained were summed. Because the voice is a single physiological parameter in contrast to the three measured by the polygraph, the cut-off points between inconclusives and decisions were modified. With the voice, a score of +3 or higher resulted in a conclusion that the suspect had told the truth; between + or -2, inclusive, was inconclusive; and with a score of -3 or lower it was concluded that the suspect had lied on the test.

Using the cut-off points of + or - 3 in order to make a definite conclusion, 6 of the 14 PSE analyses were inconclusive. Of the 8 decisions that were made, all agreed with the decisions made on the basis of the polygraph. Using a normal approximation to the binomial model, (8, ½), it was found that this was significant ($p < .01$).

Because each polygraph parameter was scored individually at the time the polygraph charts were evaluated, it is possible to list the raw scores for each of the 3 polygraph measures plus the voice. These scores are shown in Table 5.

By taking the total polygraph score as being the criterion it is possible to rank the four physiological parameters in the order of their agreement with the total polygraph score. Since the voice analysis was completed some time after the polygraph examination, the voice score did not contribute to the polygraph score. A rank of 1 indicates that the parameter was the most effective one with that particular individual; a rank of 4 indicates that it was the least helpful. Table 6 shows the ranks for each component with each suspect. Brackets around a rank indicate that that component disagreed with the total polygraph score, i.e., had that component been used alone in the absence of the other components, the examiner would have made a different decision.

When the ranks were averaged for each component over all 14 suspects, it was found that the Skin Resistance Response was the most effective single component with a mean rank of 2.46, followed by respiration and the cardio with mean ranks of 2.71 and 2.89, respectively.

TABLE 5
Scores of each individual component on each criminal suspect

Subject	Respiration	SRR	Cardiovascular	Voice
1	0	-9	-2	-4
2	0	-2	-6	-3
3	0	-6	-1	-10
4	0	-12	+3	-5
5	-13	-8	+1	-2
6	-1	-3	-2	-4
7	0	-8	-6	-4
8	-4	-15	+4	-2
9	-16	-5	-1	-1
10	+1	-7	0	+1
11	-6	-3	+1	-4
12	-11	-13	0	0
13	-5	-16	-5	-4
14	-4	-11	0	-1

The selection of a cut-off point of + or - 3 before making a definite decision when using a single component is somewhat arbitrary. One could argue that such a large inconclusive region is unduly conservative, that any non-zero score could be sufficient to make a decision when necessary. Table 7 shows the number of errors and the number of inconclusives for each individual component when the boundaries of the inconclusive region are decreased to scores of zero. It is immediately obvious that the Skin Resistance Response was by far the strongest single parameter, and that the cardiovascular measure was of very little help; it never reached statistical significance. This latter finding was unexpected, since previous research was found cardiovascular responses to be of use (Barland, 1972a; Kugelmass & Lieblich, 1966; Violante and Ross, 1964). The lack of significance here probably resulted from the small sample size.

Because numerical evaluations were made of all responses, it is of interest not only to examine the ranks of the various parameters, but also the correlation between them. The Pearson product-moment correlation between the PSE scores and the composite polygraph scores was -.359, which was not significant.

In view of the fact that the PSE analyses in this study had all been made after the polygraph examinations had been completed, it is possible that the

(6.25%). Chi square analysis (Siegel, 1956) indicated that this result was not significantly below chance (p > .10). Combining first and second guesses for each S, the analyst was right three times (18.75%) where chance would be 6.4 (40%). Chi square analyses of the frequencies of all three sets of numbers in Table 1 indicated no significant bias in the numbers chosen by the Ss and E (p > .30).

In arriving at a decision, the E analyzed the 9 PSE charts both individually and collectively, as illustrated by the matrix from Subject 14 in Table 2. This type of matrix was used because it could provide information concerning the most accurate PSE mode and the most effective type of psychological precursor for optimum detection: situational stress only, a wager, or verbal psyching. S 14 was the one subject whose lie was correctly detected on the analyst's first guess. Inspection of the matrix in Table 2 shows that the decision was relatively easy to make. Several of the matrix squares have two choices listed. This procedure was used by the E to help him make the final decision based upon the individual charts.

TABLE 2
Score Matrix, Subject 14, low-stress experiment

PSE	Chart 1	Chart 2	Chart 3	Overall
Mode I	32 or 34	32	32 or 21	32
Mode III	33 or 35	33	32	33
Mode IV	35 or 32	32 or 34	32	32
Overall	32	32	32	32 or 33

However, with most Ss, there was remarkably little consistency in stress patterns from one trial to the next, nor even among the three modes within a single trial. A more typical matrix is shown in Table 3.

TABLE 3
Score Matrix, Subject 6, low-stress experiment

PSE	Chart 1	Chart 2	Chart 3	Overall
Mode I	33 or 35	35 or 31	34 or 32	35
Mode III	35 or 32	31 or 33	33	33
Mode IV	35 or 33	33	33 or 31	33
Overall	35	31	33	33 or 35[a]

[a]The number Subject 6 had actually lied about was number 31.

It is easy to hypothesize that inconsistencies from one trial to another could be caused by changes in the S's psychological set. There was relatively little stress involved in his lie, so his attention was not steadfastly focused upon his lie throughout the experiment. However, the reason for the lack of consistency from one mode to the next within a single trail is disturbing. Several technical differences between the various modes may have contributed to the inconsistency. Mode I differs from Mode III in two major operational respects. First, the tape playback speed differs with using the two modes. This results in different frequencies being fed into the PSE. Second, it is customary to keep the playback volume constant during Mode I analysis, but to vary the playback volume during a Mode III analysis in order to obtain the optimum amplitude of wave form for visual inspection. Thus, Mode I is more sensitive to variations in the level of loudness of the speaker's voice than is the Mode III analysis. Mode IV differs from the other modes in that the low volume portion of the signal is enhanced and the high volume portion of the signal is suppressed, resulting in a more even visual, pattern. Because of the different form of the pattern, it may be that criteria of stress used with Mode III do not fully apply to Mode IV though the manufacturer's hypothesis concerning the physiological basis of the technique implies that Mode III and IV stress criteria should be the same.

Because of the very small number of hits, it was impossible to find any significant difference in accuracy between the three modes. Each mode, evaluated independently of the other modes, was correct one time out of 16 based upon the analyst's first guess. Likewise, there was no significant difference between the three trials with each S, based upon a global evaluation of all three components on each trial. As is shown in Table 4, the analyst had two hits when he evaluated the first trial by it itself, five hits when he evaluated the second trial by itself, and two hits when he evaluated the third trial alone. Because of the inconsistencies from one trial to the next, he had only one hit when he looked at all three trials together. Normally, the more data the analyst has available, the more accurate his decisions tend to be. In this case the opposite was true.

Table 4 also shows the number of hits when each PSE mode was evaluated separately on each trial. It can be seen that if the analyst had based his decision solely upon Mode III on the second trial, he would have been right 6 times out of 16, which approaches significance using a normal approximation to the binomial mode (16, 1/5, (p < .10). One more hit would have been significant. It would be incorrect to attach any importance to the near-significance of this one matrix square, for with sixteen squares one would expect at least one square to show this level of significance by chance alone.

TABLE 4
Number of Hits with Each Mode on Each Trial, low-stress experiment

PSE	Trial 1	Trial 2	Trial 3	Overall	Total
Mode I	1	4	2	1	8
Mode III	2	6	2	1	11
Mode IV	4	2	3	1	10
Combined	2	5	2	1	
Total	9	17	9	4	29

Thus, on the low-stress laboratory experiment, no significant findings were obtained. This finding was unexpected in view of the findings of a prĕvious, high-stress experiment, reported here as Experiment 2.

Experiment 2 - - High Stress

Fourteen criminal suspects undergoing polygraph examinations were utilized for the high stress study. The criminal suspects were being examined as part of a major research study of the detection of deception being conducted at the University of Utah. The criminal suspects were referred for examination by various police departments, prosecutors, and defense attorneys in Utah and Nevada. The subjects ranged in age from 18 to 37 with a mean of 27.9. There were 12 male and 2 female suspects. The incidents of which they were suspected included murder (2), rape (1), rape victim (1), grand larceny (2), sale of illegal drugs (4), forgery (1), reckless driving (1), armed robbery (1), and improper police conduct (1). The educational level ranged from 8 to 15 years with a mean of 11.4 years of formal education. All suspects were examined on field-model Keeler or Stoelting polygraphs which recorded respiration, the Skin Resistance Response, and cardiovascular

Use of Voice Changes in the Detection of Deception

Gordon H. Barland

It has long been known that short-term physiological changes occurring in persons can be highly accurate in determining whether that person is telling the truth or not, provided that the proper physiological parameters are monitored under adequately controlled conditions. This finding has been supported in both experimental laboratory studies and in examinations of criminal suspects. The most frequently used parameters include the Skin Resistence Response, respiration, and cardiovascular activity, but numerous other parameters can also be used (Barland & Raskin, 1973; Orne, Thackray, & Paskewitz, 1972). Recently, a new technique has been developed which is believed to detect short-term physiological changes that occur in the voice when a person is under stress, as when he is lying.

The use of the voice to assess short-term changes in the level of stress of an individual, would offer a number of advantages over current psychophysiological monitoring methodology. Because no sensor need be attached to the subject, there would be no discomfort to the subject. The subject would be free to move around and would not necessarily be aware that he was being monitored. This would reduce the amount of situational stress which may confound the interpretation of certain types of studies. Use of the voice to measure stress would also permit the acquisition of data, under certain circumstances, by an observer remote in distance or remote in time. The Soviets have monitored voice stress levels of cosmonauts during space flights (Simonov & Frolov, 1973), O'Toole (1973) used voice stress analysis in an investigation of the assassination of President Kennedy.

This paper describes the results of two experiments assessing the validity of voice stress analysis for the detection of deception. The first experiment was a low-stress experiment of detection of deception in a controlled, laboratory situation. The second experiment was of criminal suspects undergoing polygraph examinations in which their verbal answers to the test questions were tape recorded. The latter situation was thus a high-stress one in which the results of the voice analysis were compared with the autonomic responses recorded by the polygraph.

Experiment 1 - - Low Stress

Sixteen students (14 male, 2 female) taking an undergraduate psychology course in the detection of deception at the University of Utah volunteered for this study. The subject (Ss) appeared one at a time and were asked to choose one of five numbers ranging from 31 through 35. They were then instructed to write their choice on a 3 x 5 inch card and to pin the card up facing them so that they could see the number plainly, but the experimenter (E) could not. At no time during the testing and decision-making portions of the experiment did the E know what number had been chosen by the S.

The S was then given a routine peak of tension test (Barland & Raskin, 1973) to determine what number he had picked. The S was told that he would be asked nine questions concerning the number he had written on the card, and he was to answer all questions "no." The questions were: "Regarding the number that you wrote on that card, is it the number 29?" "Is it the number 30?", and so on, in sequence, through 37. The first two and last two numbers were inserted in order to absorb the initial orienting response and to serve as anchors for the peak.

Previous research has shown that the more emotionally involved a S is with his lie, the more easily it is detected (Gustafson & Orne, 1963). Therefore, in order to increase the emotional involvement of the S, he was asked after the first trial if he would like to try the test again, this time with a 50¢ wager. It was explained that, when the tape recording was analyzed, if the analyst was correctly able to identify the number which the S had written on the card, the S would pay E 50¢. The analyst would also make a second guess. If the second guess was correct, neither the S nor the E would pay any money. However, if the S had picked any of the three remaining, unguess numbers, then the E would pay the S 50¢. Fifteen of the 16 Ss accepted this wager. All 16 Ss were then asked the same questions again, in the same sequence. Following the second trial, all Ss were "psyched up" by the E in order to further increase their emotional involvement. This was done by questioning the S's sense of morality concerning the ethics of lying. The questions were then asked a third time, this time in reverse sequence. The reason for reversing the sequence was to better differentiate the point of deception in those cases where the S may have responded ambiguously on the first two trials.

After the third trial the S identified his numbered card by signing it. The S put it into an opaque envelope, sealed it, and gave it to a neutral E who served as scorekeeper. As the analyst made his decisions as to which number the S had picked, he gave his first and second choices to the scorekeeper who compared them with the number the S had actually picked.

The Ss answers were recorded on an Uher 4000 Report-IC monophonic tape recorder at 7½ ips by means of a Realistic carboid lavalier microphone, model MC-1000, worn by the S. After the S had been dismissed, the tape was played back through a Psychological Stress Evaluator, model PSE-1, in order to convert the audio signal to a visual chart for analysis. Three analyses were made of each trial: a Mode I analysis at a tape playback speed of 7½ ips and Mode III and Mode IV analyses at a tape playback speed of 1-7/8 ips. No Mode II analysis was made because this mode is seldom used by most PSE users. There were thus 9 PSE charts available on each S (3 modes x 3 trials) at the time the E made the decision as to which number a S had written on the card. Table 1 lists the numbers chosen by each S and the two guesses made by the E for each S. The asterisks indicate the hits made by the E.

TABLE 1
Actual numbers selected by Ss and E, low-stress experiment

	Numbers picked by		
S	S	E (1st choice)	E (2nd choice)
1	32	33	34
2	32	33	35
3	32	35	32*
4	33	31	32
5	32	34	31
6	31	33	35
7	31	35	34
8	31	35	33
9	34	31	33
10	33	32	31
11	34	33	35
12	35	33	31
13	31	34	31*
14	32	32*	33
15	33	31	35
16	35	33	34

*hits by E

The results were unimpressive. In a situation where the lie is restricted to one of only five possibilities, with an N of 16, chance detection would be 3.2 hits (20%). However, after analysis of the voice, the analyst made only one correct decision on his first choice

Michigan State University, East Lansing, Michigan 48824.

Reference Notes

1. Dektor Counterintelligence and Security, Inc., PSE orientation course manual. Unpublished undated material. (Available from Dektor CI/S, Inc., 5508 Port Royal Road, Springfield, Virginia 22151.)

2. Kubis, J. F. Comparison of voice analysis and Polygraph as lie detection Procedures (Report of contract DAADO5-72-C-0217 prepared for tbe U.S. Army Land Warfare Laboratory) Aberdeen Proving Ground, Md.: VS. Army Land Warfare Laboratory, August, 1973.

3. Shipp, T., & McGlone, R. Physiologic correlates of acoustic correlates of psychological stress. Paper presented at the meeting of the Acoustical Society of America, Los Angeles, November 1973.

4. McGlone, R., & Hollien, H. Partial analysis of acoustic signal of stressed and unstressed speech Proceedings, 1976 Carnahan Conference on Crime Countermeasures (BU No. 110). Lexington: ORES Publications, College of Engineering, University of Kentucky, 1976.

References

Barland, G. H. Detection of deception in criminal suspects: A field validation study. Unpublished doctoral dissertation. University of Utah, 1975

Barland, G. H., & Raskin, D. C. Detection of deception. In W. F. Prokasy & D. C. Raskin (Eds.), Electrodermal activity in psychoiogical research. New York: Academic Press, 1973.

Fay, P., & Middleton, W. The ability to judge truth-telling or lying from the voice as transmitted over a public address system. *Journal of General Psychology,* 1941, *24,* 211-215.

Gustafson, L. A., & Orne, M. T. The effects of heightened motivation on tbe detection of deception. *Journal of Applied Psychology,* 1963, *47,* 408-411.

Gustafson, L. A., & Orne, M. T. The effects of task and method of stimulus presentation on the detection of deception. *Journal of Applied Psychology,* 1964, *48,* 383-387.

Horvath, F. The effect of selected variables on interpretation of polygraph records. *Journal of Applied Psychology,* 1977, *62,* 127-136,

Kugelmass, S., & Lieblich, I. Effect of realistic stress and procedural interference in experimental lie detection. *Journal of Applied Psychology,* 1966, *50,* 211-216.

Kugelmass, S., Lieblich, I., Ben-Ishai, A., Opatowski, A., & Kaplan, M. Experimental evaluation of galvanic skin response and blood pressure change indices during criminal interrogation, *Journal of Criminal Law, Criminology, and Police Science,* 1968, *59,* 632-635.

Lieblich, I., Naftali, G., Shmueli, J., & Kugelmass, S. Efficiency of GSR detection of information with repeated presentations of series of stimuli in two motivational states. *Journal of Applied Psychology,* 1974, *59,* 113-115.

Lippold, O. Physiological tremor. *Scientific American,* 1971, *224,* 65-73.

Olechowski, R. [Experiments on voice modulation while telling lies.] *Zeitschrift Fur Experimentelle Und Angewandte Psychologie,* 1967, *14(3),* 474-482. (Psychological Abstracts, 1968, *42,* 484.)

Orne, M. Implications of laboratory research for the detection of deception. *Polygraph,* 1973, *2,* 169-199.

Reid, J. E., & Inbau, F. E. *Truth and deception. The polygraph ("lie detector") technique* (2nd ed.). Baltimore, Md.: Williams & Wilkins, 1977.

Williams, C., & Stevens, K. Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of Amercia,* 1972, *52,* 1238-1250.

Yankee, W. J. An investigation of sphygmomanometer discomfort thresholds in polygraph examinations. *Police,* 1965, *9,* 12-18.

Table 3
Distribution of Evaluators' Ranks of the Responses to All Critical Items and All Noncritical Items Averaged Across Subjects' Two Trials

	PSE		GSR	
Mean rank	Critical items	Non-critical items	Critical items	Non-critical items
1.00		2	11	
1.25		3	2	2
1.50	1	7	7	6
1.75	3	21	1	4
2.00	1	10	3	6
2.25	6	23	2	6
2.50	4	22	6	17
2.75	7	17	2	13
3.00	6	22	2	19
3.25	8	24	3	11
3.50	4	17		16
3.75	5	29	1	16
4.00	4	17		16
4.25	1	14		4
4.50	3	7		12
4.75		2		9
5.00	1	3		3

Note. PSE = Psychological Stress Evaluator; GSR = galvanic skin response.

Discussion

These results are remarkably consistent with those reported by Kubis (Note 2). On the one hand, PSE analysis yielded hit rates only at chance levels. On the other hand, the hit rates obtained in GSR analysis were far superior to those obtained in PSE analysis, and overall, well beyond chance levels.

The low detection efficiency in PSE analysis precluded the discovery of any significant effects for any of the independent variables examined. However, both the use of the polygraphs' blood pressure cuff and repeated trials did affect GSR analysis. Kugelmass and Lieblich (1966) have found that in low-risk situations the blood pressure cuff tends to reduce the contrast between responses to relevant and nonrelevant options, that is, it appears to lower the signal-to-noise ratio. Their findings are supported by the results in this research. But, it is important to point out that the interference effect of the blood pressure cuff on GSR responses appears to diminish with increasing levels of stress (Kugelmass, Lieblich, Ben-Ishai, Opatowski, & Kaplan, 1968), and there is a growing body of evidence showing that the detection efficiency of the GSR in real-life situations is not substantially affected by the blood pressure cuff (Barland, 1975; Barland & Raskin, 1973).

Generally, hit rates observed in GSR analysis in this research were quite consistent with those reported in previous experimental research using the guilty-information paradigm (Gustafson & Orne, 1963; Gustafson & Orne, 1964; Kugelmass & Lieblich, 1966; Lieblich, Naftali, Shmueli, & Kugelmass, 1974). However, in the present study, unlike most prior research, hit rates were calculated separately for each of two consecutive trials;

thus, it was possible to observe a systematic difference between those trials in regard to their detection efficiency. That difference obtained whether GSR responses were subjectively or objectively scored. For instance, analysis of variance carried out only on objectively assigned dichotomous scores (a rank of 1 being a correct detection, all other ranks incorrect) showed a significantly greater detection rate for Trial 1 than for Trial 2, 67.5% to 42.5%, $F(1,36) = 5.06$, $p< .03$.

The standard procedure to minimize the confounding effects of repeated testing is to average responses across trials on an intra-subject basis. Such a procedure generally results in higher detection rates (Lieblich et al., 1974). Unfortunately, because PSE response data are not readily objectively scored, it was not feasible to carry out such a calculation on both PSE and GSR data in the present research; therefore, the difference between this and other research in the manner in which hit rates were calculated justifies some caution in directly comparing results. Nevertheless, the marked similarity between this and other experimental research regarding GSR analysis suggests, as Kugelmass et al. (1968) and Orne (1973) have also reported, that the difference between field and laboratory equipment probably does not explain the general disinclination of field examiners to rely on GSR data (Horvath, 1977; Reid & Inbau, 1977).

There are two limitations in this study pertaining to the results of PSE analysis that deserve brief mention. First, the full technical capability of the PSE was not evaluated. Subjects' vocal responses were monosyllabic in nature and were analyzed in only one of the four display modes of the PSE. Second, this study did not involve any overt manipulation of the subjects' motivational level. It has been demonstrated by Gustafson and Orne (1963) that the detection efficiency of the guilty-information paradigm, at least with respect to measures of electrodermal activity, depends to a considerable degree on subjects' motivation to deceive. It is not known if an increase in motivation increases detectability with the PSE or if a certain degree of psychological stress, not achieved in the present study, is necessary to maximize the effectiveness of the PSE. In spite of those possibilities Barland's (1975) findings in actual criminal suspects, who are presumed to be highly motivated to deceive, suggest that even in such circumstances the PSE is not effective in detecting deception.

The detection rates in PSE analysis in this study were not dissimilar to those reported by other investigators who made non-instrumental attempts to detect deception in the human voice (Fay & Middleton, 1941; Olechowski, 1967). Although electronic analysis of the speech spectrum would appear to be the more reliable of the two procedures, the acute inter- and intrasubject variability in the voice, and the lack of an adequate specification of the precise relationship between the components of the voice spectrum and emotional states (Williams & Stevens, 1972), present complex and formidable problems in using the voice to detect deception. In fact, contrary to the relationship claimed to exist between emotional stress and low frequency tremors in the voice, Shipp and McGlone (Note 3) found no electromyographic evidence of such tremors in the laryngeal muscles in vocalization of truthful or deceptive utterances. Similarly, McGlone and Hollien (Note 4), who spectrographically analyzed speech samples of subjects who read a passage in an unstressed condition and those of subjects who read a passage while receiving a series of electrical shocks, found no low-frequency energy in the speech samples of either group of subjects. Thus, neither the PSE nor its theoretical premise appear to be useful approaches to resolving the problems associated with detecting stress in the voice. More specifically, as a means of detecting deception, at least within the constraints of this experimental setting, the PSE was highly unreliable and was clearly much less useful than the traditional field measure of electrodermal activity.

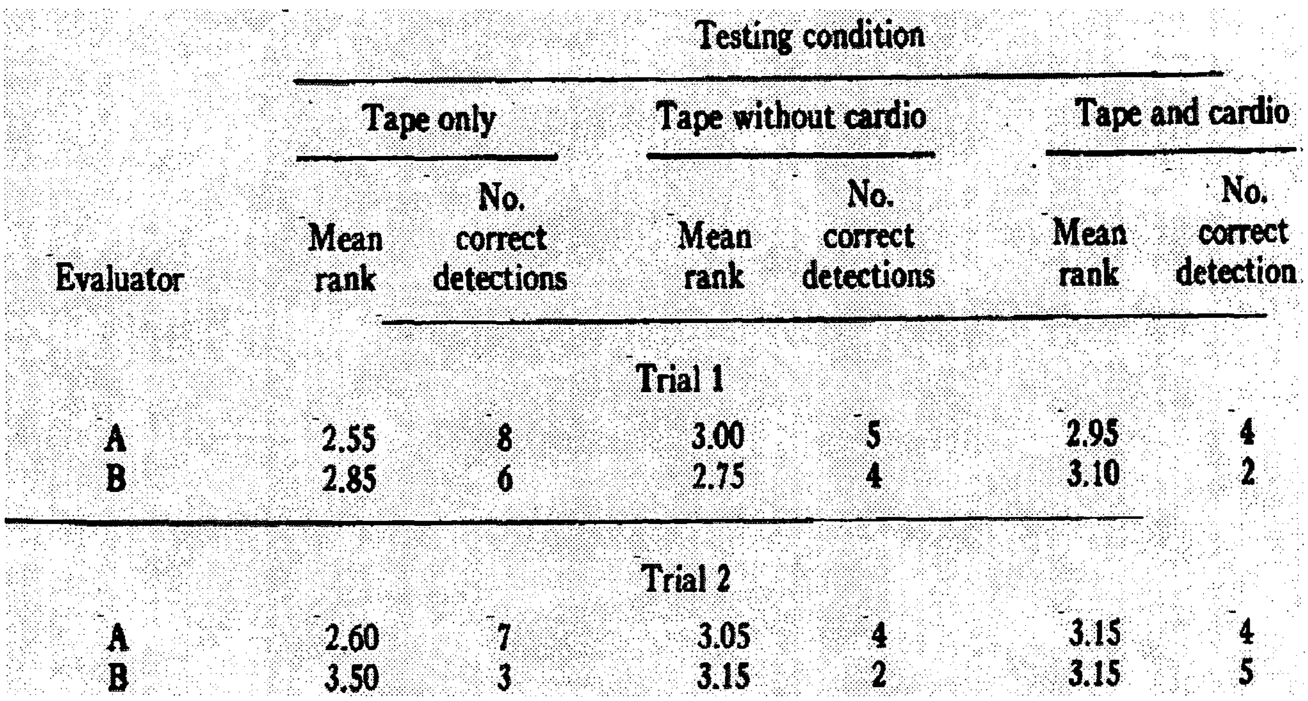
Acknowledgements

The author gratefully acknowledges the assistance of Earl W. James, Howard Timm, and Steven Bagnasco in the collection of the data, and the cooperation of Alan Bell, Edward Kupec, and Gil Gray, all of Dektor CI/S.

Requests for reprints should be sent to Frank Horvath, School of Criminal Justice,

was not significant, *F* (1, 36) = 3.19, p < .08, but the effect for trials was *F* (l, 36) = 6.10, p < .02, the average detection rates being 68.8% in Trial 1 and 42.5% in Trial 2.

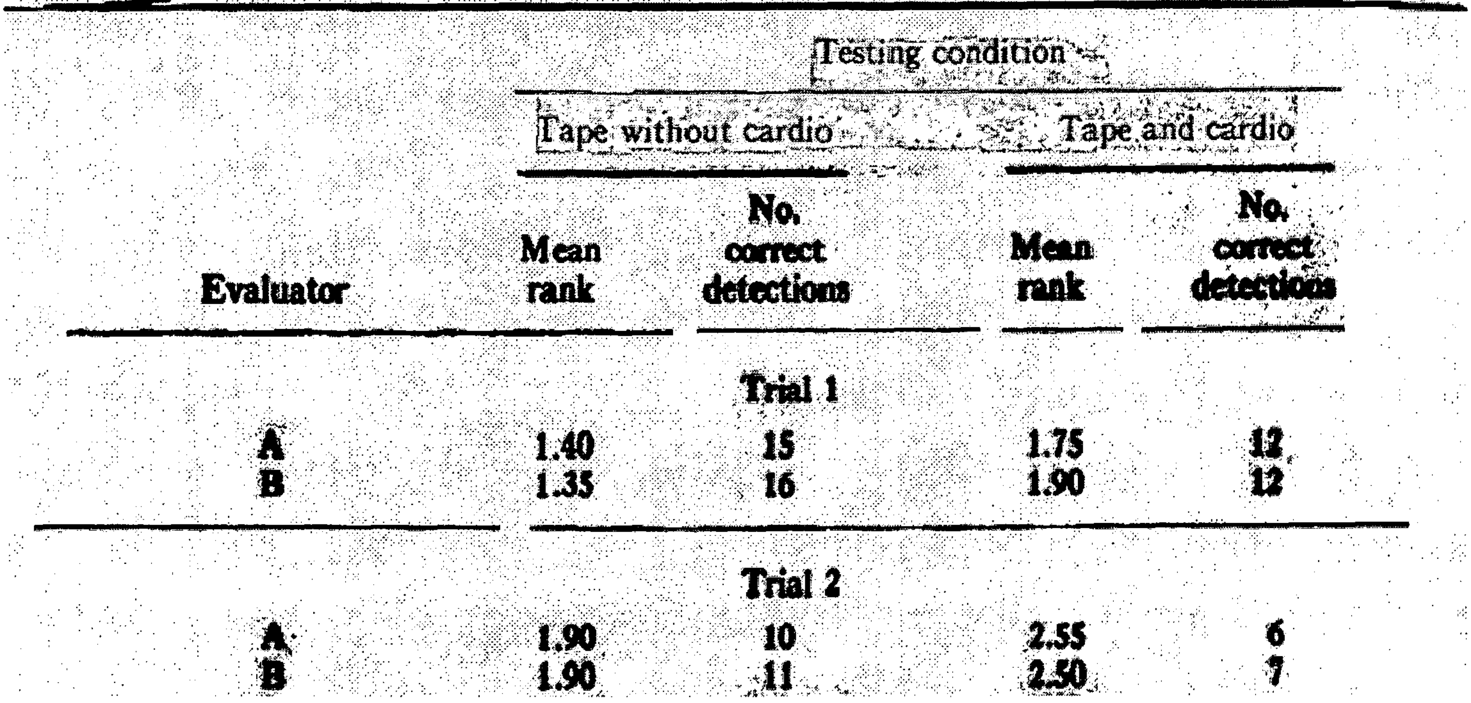
Table 1
Evaluators' Mean Ranks to Critical Items and Number of Correct Detections in Psychological Stress Evaluator Analysis



	Testing condition					
	Tape only		Tape without cardio		Tape and cardio	
Evaluator	Mean rank	No. correct detections	Mean rank	No. correct detections	Mean rank	No. correct detection
			Trial 1			
A	2.55	8	3.00	5	2.95	4
B	2.85	6	2.75	4	3.10	2
			Trial 2			
A	2.60	7	3.05	4	3.15	4
B	3.50	3	3.15	2	3.15	5

Note. Each evaluator made 20 calls in each testing condition, each trial being independently analyzed. Using the binomial distribution (n = 20, 1/5), a result of eight or more hits is significantly (p < .03) greater than chance expectancy.

Table 2
Evaluaton' Mean Ranks to Critical Items and Number of Correct Detections in Galvanic Skin Response Analysis



	Testing condition			
	Tape without cardio		Tape and cardio	
Evaluator	Mean rank	No. correct detections	Mean rank	No. correct detections
		Trial 1		
A	1.40	15	1.75	12
B	1.35	16	1.90	12
		Trial 2		
A	1.90	10	2.55	6
B	1.90	11	2.50	7

Note. Each evaluator made 20 calls in each testing condition, each trial being independently analyzed. Using the binomial distribution (n = 20, 1/5), a result of eight or more hits is significantly (p < .03) greater than chance expectancy.

Each evaluator, however, did obtain an overall detection rate in each trial significantly greater than chance expectation (chi-square).

Analysis was also carried out on the raw GSR ranks assigned by evaluators to the critical items; smaller mean ranks indicated greater efficiency in detection. That analysis revealed that evaluators' mean rank in Trial 1, 1.60, was significantly lower than that in Trial 2, 2.21, *F*(1,36) = 4.52, p< .04; and that the mean rank assigned to GSR responses recorded without an operational blood pressure cuff, 1.64, was significantly lower than the mean rank assigned when the cuff was inflated, 2.18, *F*(1,36) = 4.28, p < .04.

Table 3 displays the distribution of evaluators' mean ranks, calculated by averaging ranks across each subject's two trials, to all critical items and all noncritical items in both PSE and GSR evaluation. Assuming any fixed cutoff point on the mean rank dimension shows the relatively greater detection efficiency of the GSR compared to that of the PSE. A cutoff of 2.25, for example, yields a proportion of 17:60 hits (critical items ranked at or less than the cutoff) and a proportion of 66:240 false alarms (noncritical items ranked at or less than the cutoff) for the PSE. At that same cutoff point, the GSR yields a 26:40 hit rate and a 24:160 false-alarm rate.

carried out the testing in a small, quiet, private office. The assistant initially conducted an interview lasting about 30 min during which he gathered brief background information, explained the nature of the testing apparatus, and the theory of detection of deception. To those subjects who were assigned to the two testing conditions in which the polygraph instrument was to be used, he gave a short demonstration of that apparatus. He then explained the testing procedure, and when assured that each subject understood the procedure, he operationalized the appropriate apparatus and carried out the testing.

The testing procedure, which was identical for all subjects except for the apparatus used, consisted of presenting to each subject a deck of five numbered cards face down. The subject chose one of the cards, looked at the number on it, and then, out of view of the assistant, wrote the number and his name on 3 small slip of paper; he then placed both the card and the paper slip face down in front of him. At no time prior to the completion of the testing was the assistant aware of the card number a subject had chosen.

The testing consisted of asking the basic question "Did you pick card number __?" in two consecutive continuous trials. The subject was instructed to answer no to each card number during each trial and to sit motionless with his eyes closed throughout the testing. In the first trial the card numbers were called in ascending sequence, preceded and followed by a buffer number, that is, a number known not to be in the deck. Immediately following the second buffer item the subject was asked a pivotal question, "Is your first name ____ ?", to which a yes response was required. A second trial was then conducted; in this trial the card numbers asked in the first trial were called in reverse order. During both trials, card numbers were called at about 20-sec intervals. All subjects had advance knowledge that in the first trial card numbers were to be called in ascending sequence; in the second, descending. The numbers, however, were not consecutive, and subjects were aware only of the number on their chosen card.

Upon completion of the testing, the assistant noted on the polygraph charts, where appropriate, and on the tape recording an identification code number for each subject. Then, the polygraph charts were prepared for evaluation by cutting each subject's charts into two halves, one half consisting of Trial 1, one half of Trial 2; each half was then coded in such a manner that the two halves could not be matched without knowledge of the coding scheme.

From the tape recordings, PSE charts were made by charting each subject's no responses to the card options separately for Trial 1 and Trial 2. The charts for each trial were then coded in a manner to prevent matching. All PSE charts were made on a PSE-101 in Mode 3 at a constant speed reduction of 4:1; that is, PSE charts were produced by playing back subjects' verbal responses at 1 7/8 ips.

Two trained and experienced field polygraph examiners, both also having been trained in the use of the PSE by the manufacturer, independently and subjectively evaluated the PSE and the polygraph charts in a blind manner. In the evaluation of the PSE charts, each of the five possible options in each trial was ranked from 1 to 5, 1 being assigned to the option believed to be the chosen card, that is, the response indicating the greatest stress (least FM) according to criteria taught by the manufacturer, and 5 being assigned to the option indicating the least stress. The polygraph charts were ranked in a manner identical to that carried out on the PSE charts, except that in this case each recorded physiological measure was separately ranked. Although only the GSR rankings were analyzed, it is necessary to point out that those rankings were not necessarily independent of other polygraphically recorded data. Because of such possible contamination, GSR responses were also objectively scored. An assistant, without any prior knowledge of the experiment, ranked each GSR response in each trial for each subject by assigning a rank of 1 to the response attaining the greatest millimeters of amplitude in the period starting with stimulus onset to 15 sec following stimulus offset. The response with the second greatest amplitude was assigned a rank of 2 and so forth; to the case of ties, mean ranks were assigned.

The rank assigned by each evaluator to the card option actually chosen by each subject was determined. If the chosen card was assigned a rank of 1, it was considered a correct detection, while if it was more than 1 it was considered as incorrect. Thus each evaluator's rank on the card actually chosen by each subject was dichotomously scored, a 1 being assigned to a correct detection, a 0 to an incorrect detection. Unless specified otherwise, statistical analysis was carried out by subjecting evaluators' dichotomous scores to a four-way analysis of variance with repeated measures. The four factors were testing condition (tape, tape without cardio, tape and cardio); sex (female, male); trials (1 and 2); and evaluators (A and B). The latter two factors were treated as repeated measures. All statistical testing employed a .05 rejection region.

Results

PSE Analysis

The major findings pertaining to the PSE analysis for each evaluator are shown in Table 1, which displays, by testing condition, the mean ranks to subjects' chosen cards (critical items) and the number of correct detections in each trial; smaller mean ranks indicate greater efficiency in detection.

Each evaluator made 60 calls in each of two trials, each trial being independently considered. Application of the decision rule previously specified and disregard for the sex of the subjects and the testing conditions showed that evaluators averaged 24.2% correct calls in Trial 1; in Trial 2 20.8% of the calls were correct. The difference between trials was not significant, F(l, 54) = .25, p > .10; nor were either of the evaluators' overall hit rates in either trial significantly greater than chance expectancy of 20% (using the chi-square technique). Interevaluator agreement, determined separately for each trial by calculating Pearson's r on the ranks assigned by evaluators to the subjects' chosen cards, was .31 and .45 for Trial one and Trial two, in order. The difference in the detection rates between conditions was not significant, *F* (2, 54) = 1.79, p > .10, and there were no significant effects associated with sex or evaluators. Moreover, as indicated in Table 1, a binomial test of each evaluator's detection rate within testing conditions showed that those rates were not generally above chance levels. Similarly, analysis of variance carried out on evaluators' ranks to critical items failed to disclose any significant effects for testing conditions, *F* (2, 54) = .35, p > .10; trials, *F* (l, 54} = .96, p > .10; or for any of the other factors.

GSR Analysis

Physiological data recorded by polygraph were available, of course, in only two testing conditions; only the findings pertaining to evaluation of GSR are reported here. To determine whether evaluators' subjective judgments of GSR responses were influenced by their inspection of other polygraphically recorded data, evaluators' ranks on subjects' chosen cards were correlated with those assigned by objective measurement. Pearson's r, averaged for the two evaluators, was .76 in Trial 1 and .65 in Trial 2. However, chi-square tests did not reveal any significant differences in the detection rates obtained by objective or subjective methods. Hence, because those two methods yielded similar results and because PSE responses were not objectively scored, only the results pertaining to subjective evaluation of GSR will be reported.

Each evaluator made 40 calls in each of two trials, each trial being independently considered. There was high interevaluator agreement in ranking responses to the chosen cards, Pearson's r being .92 for both Trial 1 and Trial 2. To facilitate comparison to the PSE findings, Table 2 shows each evaluator's mean rank to chosen cards and number of correct detections in each testing condition and in each trial. In all but the "tape and cardio" condition in the second trial, each evaluator's detection rate was significantly greater than chance expectation (binomial).

Analysis of variance was carried out on evaluators' GSR detection rates; that analysis, which was identical to that previously specified except, of course, there were only two levels of testing conditions, did not reveal any significant effects associated with sex or evaluators. The effect for testing conditions

An Experimental Comparison of the Psychological Stress Evaluator and the Galvanic Skin Response in Detection of Deception

Frank Horvath

Abstract

The Psychological Stress Evaluator (PSE), which is asserted to be a voice-mediated lie detector, and the galvanic skin response (GSR), recorded with a standard field polygraph instrument, were used to detect nonrisk lies about numbered cards concealed by a sample of female (n = 30) and male (n = 30) college students. Evaluation of response data was subjectively carried out by two trained evaluators; their interrater agreement was .38 for PSE analysis and .92 for GSR evaluation. The hit rates obtained in PSE analysis were at chance levels and were not significantly affected by the sex of the subjects, simultaneous use of both PSE (tape recording) and polygraph apparatus, repeated trials of testing, or evaluator differences. Evaluations based on GSR analysis generally exceeded chance levels; however, hit rates were significantly (p < .05) higher in a first trial of testing than in a second trial. These findings were consistent with previous research and do not indicate that the PSE is effective in detecting deception.

The Psychological Stress Evaluator (PSE) is a device that is said to be useful in detecting emotional stress in the voice. According to its manufacturer, Dektor CI/S, Inc., the PSE detects inaudible and involuntary frequency modulations (FM) in the 8-12 Hz region. These frequency modulations, whose strength and pattern are inversely related to the degree of stress in a speaker, are believed to be a result of physiological tremor or microtremor (Lippold, 1971) that accompanies voluntary contraction of the striated muscles involved in vocalization. During non-stressful periods the modulations are under control of the central nervous system. As stress is imposed the autonomic nervous system gains dominance, resulting in a suppression of FM. This suppression, indicative of emotional stress, is displayed by the PSE as a characteristic blocked or rectangular wave form.

The PSE processes voice frequencies, preserved on a normal tape recording, using electronic filtering and frequency discrimination techniques. The stress-related FM patterns, displayed on a moving strip of heat sensitive paper, can be processed in four different modes of display (1-4) for either gross or more detailed analysis. And, because the recovery of the FM indicator spontaneously occurs with the removal of the stressing stimulus, stress in either narrative or monosyllabic speech can be evaluated (Dektor, Note 1).

The PSE is primarily marketed as a voice-mediated lie detector, more versatile but no less effective than the traditional polygraph instrument (Dektor, Note 1). To date, that claim has been investigated in only two scientifically acceptable studies. The most recent of these was a study carried out by Barland (1975) to determine the validity of the polygraph and the PSE in detecting deception in suspects involved in actual criminal investigations. In brief, Barland found that the accuracy of each physiological measure recorded with the polygraph instrument exceeded chance levels, whereas the accuracy of the PSE did not.

Barland's (1975) findings were essentially similar to those reported by Kubis (Note 2), who conducted an elaborate but laboratory-based study involving mock crime situations. Kubis found that the hit rate for the PSE was at chance levels, 33%; the hit rate for the polygraph was 75%, and the accuracy of judges who evaluated only the behavior of the subjects undergoing testing surpassed that obtained with the PSE. Kubis also reported, however, that the accuracy of PSE analysis on tape recordings made without the simultaneous use of polygraphic apparatus was 53%, whereas accuracy was 19% in analysis of recordings of polygraphically monitored subjects. Kubis hypothesized that the physical discomfort produced by the polygraph's blood pressure cuff, actually an occluding plethysmograph, and the absence of stresses associated with the attachment of polygraph apparatus, produced clearer voice records and thus more accurate PSE evaluations.

The purpose of the present study was to investigate the validity of the PSE in a "guilty-information" paradigm (Gustafson & Orne, 1964), and specifically, within that context, to determine if, as Kubis (Note 2) hypothesized, the simultaneous use of polygraph and tape recording apparatus reduces the effectiveness of PSE analysis. Moreover, because the physical discomfort of the polygraph's blood pressure cuff increases as a function of time (Yankee, 1965), it was expected that the validity of the PSE would decrease in a second testing period immediately following a first. The galvanic skin response (GSR) was used as the physiological measure against which the accuracy of the PSE was compared.

Method

Subjects

Sixty college students, 30 female and 30 male, were recruited for an experiment in lie detection from an introductory course in criminal justice. Upon volunteering, each student completed an informed consent form that briefly outlined the nature of the experiment and promised that each student would be awarded extra credit toward his course grade for his participation, contingent only upon maintaining a scheduled appointment and completing the task.

The age range for the female subjects was from 18 to 21 years, with a mean age of 19.2 years; for the males the age range was from 18 to 31 years, with a mean of 19.9. None of the subjects bad previously participated in a detection of deception experiment.

Procedure

Twenty subjects, 10 female and 10 male, were randomly assigned to one of three testing conditions. Subjects assigned to the "tape only" condition were tested using tape recording apparatus only. A Uher 4000 Report-IC monophonic tape recorder, operating at 7.5 in. per sec (ips), fresh l-mil polyester tape, and a Sony omnidirectional microphone, positioned in front of the subject, were used for recording. In the remaining two conditions, testing was carried out simultaneously using tape recording and polygraph apparatus. The polygraph was a standard Stoelting field instrument, recording respiration, GSR, and cardiovascular activity. Respiration was recorded by a pneumatic tube positioned on the abdomen near the level of the diaphragm, adjusted to provide a pen excursion of 1-3 cm. GSR was recorded from two stainless-steel electrodes, attached without electrolyte to the volar surfaces of the index and fourth fingers of subjects' left hand; in all cases GSR was recorded in the automatic centering mode; that mode employs a short-time constant measurement technique that eliminates information concerning response recovery time. Cardiovascular activity was recorded by an occlusive blood pressure cuff located on the upper part of subjects' right arm. The cuff was inflated to a pressure of about 90-mm Hg to record cardiovascular activity in a manner consistent with standard field practice (Reid & Inbau, 1977).

In the "tape without cardio" condition, the polygraph's blood pressure cuff was attached to the subject but was not inflated; hence, for those subjects who were assigned to that condition no discomfort was produced by the cuff and no cardiovascular activity was recorded. Subjects who were assigned to the "tape and cardio" condition were tested with a fully operational polygraph, recording the three physiological measures as previously described.

Upon reporting for the experiment, each subject was met by an assistant who

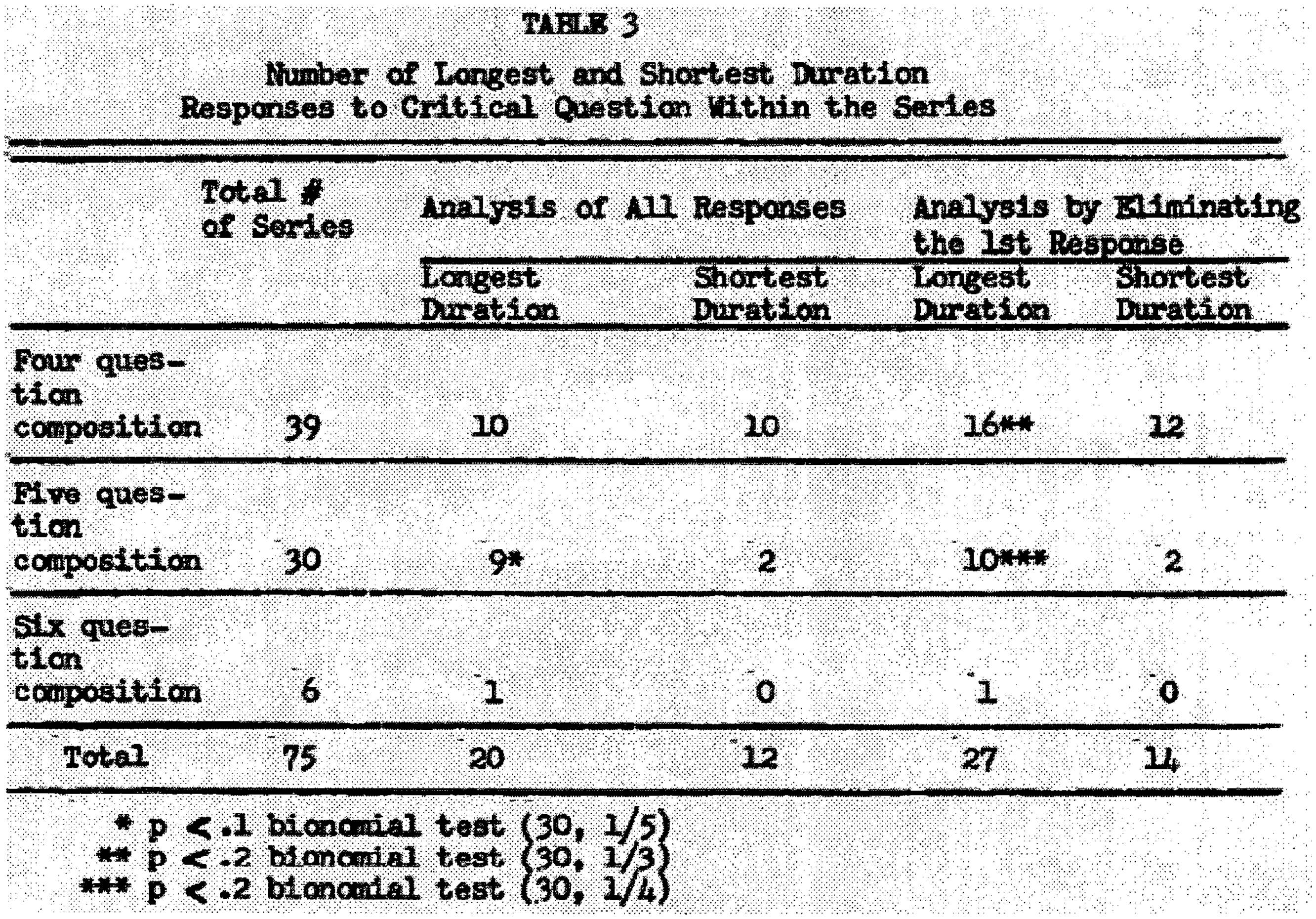


TABLE 3

Number of Longest and Shortest Duration Responses to Critical Question Within the Series

	Total # of Series	Analysis of All Responses		Analysis by Eliminating the 1st Response	
		Longest Duration	Shortest Duration	Longest Duration	Shortest Duration
Four question composition	39	10	10	16**	12
Five question composition	30	9*	2	10***	2
Six question composition	6	1	0	1	0
Total	75	20	12	27	14

* p <.1 bionomial test (30, 1/5)
** p <.2 bionomial test (30, 1/3)
*** p <.2 bionomial test (30, 1/4)

Table 3 shows the results of analysis on the duration of subjects' answers. An analysis of 5 question composition approached significance (p < .1, 30, 1/5) and an analysis of 4 and 5 question composition by eliminating the answers to the first question also approached significance (p < .2, 39; p < .2, 1/3, 30, 1/4). Both showed a tendency of longer duration for answering critical questions but this was not significant. Nevertheless, it showed a higher detection rate than by a pitch or intensity method, but it is still under 50% indicating that it is not applicable in actual cases.

In the analysis of sonagram for subject A, all 3 voice identification specialists failed to determine the answer to a critical question. Specialist (l) achieved 4/15 (26.6%) as correct decisions; the other two specialists did 2/15 (13.3%) as correct judgements. All three reported that they could not determine the deceptive answers and notable changes. Therefore, the sonagram can be judged as difficult to analyze and not reliable or adaptable for actual cases. From the results of these analyses, using pitch, intensity and duration of voices as a means to detect deception, the utility appears slim at this stage.

Acknowledgements

Suzuki and Watanabe are the Research Psychologists for the NIPS. Takeno, Kosugi, and Kasuya are the Police Psychologists for Criminal Investigation Lab., Metropolitan Police Department.

Reprinted with the permission of the authors and the NIPS from Reports of the National Institute of Police Science, 26, (1973): 62-66.

References

Alpert, H., Kurtzberg, R.L., & Friedhoff, A.J. Transient voice changes associated with emotional stimuli. *Archives of General Psychiatry*, 1963, 8, 362-365.

Fay, P. J., & Middleton, W. C. The ability to judge truth telling, or lying, from the voice as transmitted over a public address system. *Journal of General Psychology*, 1941, 24, 211-215.

Maki, M. Voice changes in critical and control answers during peak of
tension test. Research Material No. 39, Polygraph Reports, National Institute of Police Science, 1968, 92-97. [in Japanese.]

[Translation received May 12, 1979; Revised September 28, 1979; accepted for publication December 13, 1979.]

showing the intensity to each question in 1st series was traced on a paper to superimpose, but no peculiar answering characteristics to a critical question was noted. So, only the maximum intensity points of each answer were extracted and measured.

(4) A sonagram is used in voice identification. The ordinate axis gives the time and the abscissa axis indicates the frequency. The density of pattern inscribed on the graph shows the intensity of the voice frequency component. In the graph, voice duration, formant (phonetic) voice intensity and consonant are displayed, but because of the consolidation of multi-dimentional analysis, a subjective judgement was assigned to specialists. Sonagrams which analyzed the answers of subjects of each series were mounted on a board and we had three specialists on voice identification analyze them. The following instruction was given: "These charts are sonagrams of subjects' answers to questions used in the polygraph test. Questions consisted of 4 or 6 in each series. A chart of each series contains one deceptive answer. Please select one chart which shows peculiarity from others and record the number of the chart. When a judgement cannot be made, indicate this by writing so."

Results and Discussion

The difference in the pitch, intensity and duration of the voice between deception and truth has not been clear. It is not known whether pitch increases or decreases when a deceptive answer is given. A frequency of the critical question which had either the highest or the lowest pitch in each series question was counted. In an actual examination, an orienting response occurs at the first question of each series. It is not known whether this orienting response occurs in the voice pitch; therefore, after excluding the answer to the first question, a frequency indicating highest or lowest pitch in critical questioning was also counted. The results are as shown in Table 1. Among the 75 series, 10 highest (14.6%) and 13 lowest (17.3%) pitch responses in series were associated with the critical questions. A chance detection rate would be 1/2 or 50%, but the rate from the results obtained here is lower.

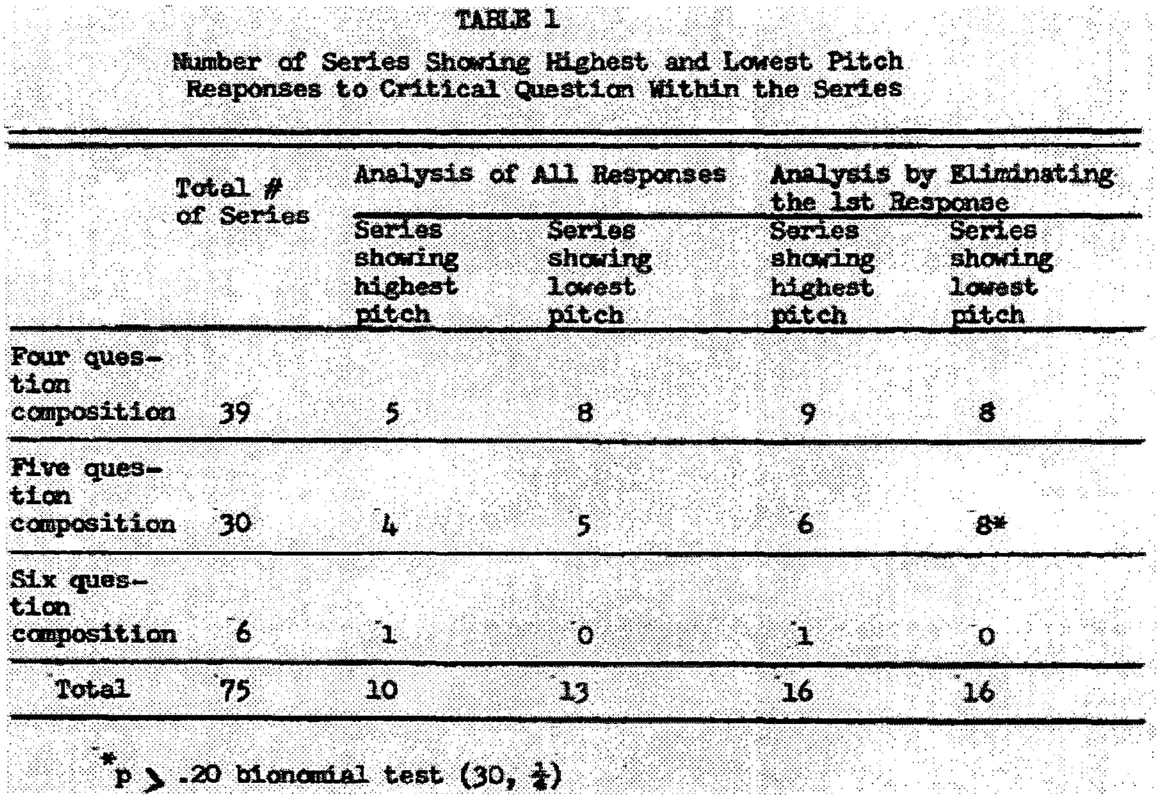


TABLE 1

Number of Series Showing Highest and Lowest Pitch Responses to Critical Question Within the Series

	Total # of Series	Analysis of All Responses		Analysis by Eliminating the 1st Response	
		Series showing highest pitch	Series showing lowest pitch	Series showing highest pitch	Series showing lowest pitch
Four question composition	39	5	8	9	8
Five question composition	30	4	5	6	8*
Six question composition	6	1	0	1	0
Total	75	10	13	16	16

*$p > .20$ bionomial test (30, ½)

If the voice pitch increases when a deceptive answer is given, the following can be assumed. Performing "m" series of question lists which consisted of "n" questions, the number of the critical responses indicating the highest pitch in a series would exceed m/n. This also applies when the pitch is lower. For example, assuming that a 4 question chart in 40 series given a ranking of 1, 2, 3, 4 from the highest pitch in each series, the count is made on the frequency of (1) and (4) during the 40 series. If the pitch decreases during deception, the frequency of (4) should be significantly higher than 10. A bionomial test of results given in Table 1 was conducted. When the first answers were eliminated, the 4 question chart was treated as a 3 question chart, 5 as 4 and 6 as 5. The analysis of 5 question 30 series after eliminating the responses to the first question, showed that the pitch was lower than the chance probability, but was not significant (n = 30, 1/4, P > .2).

A peculiar reaction does not always occur in the field polygraph test using three indices. It is also presumed that this can be said for the voice analysis. A change in the pitch is small and does not produce a satisfactory result.

The results of the analysis of intensity are as shown in Table 2. By using the binomial test for the six question composition, the series showing the maximum intensity to critical question showed a higher frequency which approached significance (p < .2, 6, 1/6). There was no sign of increasing or decreasing of voices in intensity during the questioning.

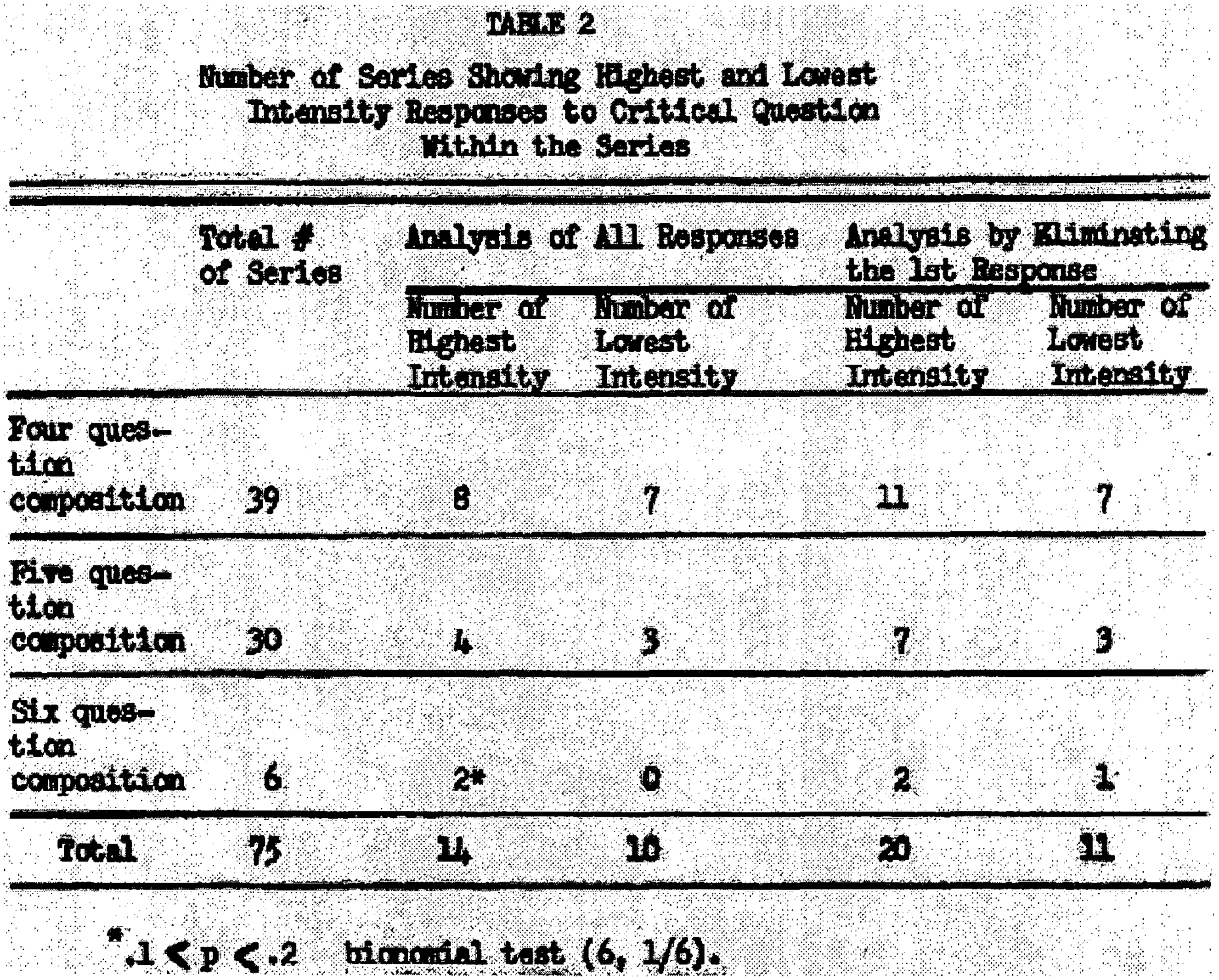


TABLE 2

Number of Series Showing Highest and Lowest Intensity Responses to Critical Question Within the Series

	Total # of Series	Analysis of All Responses		Analysis by Eliminating the 1st Response	
		Number of Highest Intensity	Number of Lowest Intensity	Number of Highest Intensity	Number of Lowest Intensity
Four question composition	39	8	7	11	7
Five question composition	30	4	3	7	3
Six question composition	6	2*	0	2	1
Total	75	14	10	20	11

*.1 < p < .2 binomial test (6, 1/6).

Possibility of Detecting Deception by Voice Analysis

Akihiro Suzuki, Shoichi Watanabe, Yutaka Taheno, Tsuneo Kosugi and Takumi Kasuya

Abstract
Measures of voice pitch, intensity, and duration were recorded and measured with apparatus used for the analysis of voice from tape recordings. Analysis was made of seventy-five answers to relevant crime questions from polygraph tests in real criminal cases in which the answers were verified as deceptive by subsequent confession or by medical jurisprudence. Each of the three methods were measured against chance, and none exceeded chance. The duration of the subject's answers showed a higher detection rate than did analysis of intensity (frequency analysis) or analysis of pitch (frequency of highest and lowest voice pitch). The authors concluded that these voice measures were not reliable or useful. [N.A.]

Preface
In order to refine the lie-detection technique, an improvement in the indices measured by the polygraph is important. Many workers have paid attention to new indices including EEG, plethysmograph, EMG and others and which restrict, to some extent, movement of the subject on account of attachment of a sensor. Since voluntary control of voice is easy, the use of the voice has not become a subject of discussion in Japan in lie detection technology. Therefore, only a few studies have been done on this subject. Maki (1968) using a noise meter studied the changes in voices; Fay and Middleton (1941) made a study on subjective analysis of voices; and Alpert et al. (1963) used two types of band-pass filters of 100-6000 Hz and 100-250 Hz to analyze voices. Maki suggested the possibility of using changes in voices as a supplemental index. Fay and Middleton showed a detection rate of 55% through use of subjective judgement and Alpert et al. showed that there was hardly any difference in truth and deception when the 100-6000 Hz filter was used but a change in voice amplitude was noted when the 100-250 Hz filter was used. Despite these effects, lie detection by voice analysis has not reached the practical stage. The voice is not only easy to record but it can be collected without the awareness of the subject being monitored. Its potential for lie detection cannot be under estimated and it should not be discounted too lightly. The key issue of voice analysis in lie detection is the method of processing information in voices. At present, a bundle of analysis methods should be studied in order to probe for the better system. The purpose of the experiment given in this report was to explore, along the above mentioned line, the possibility of lie detection by means of voice analysis.

A human voice is formed by exhalation, utterance and articulation. A sound wave passing through various parts of the vocal system (mouth, throat, lips, etc.) produces words and distinctive resonance.

If a subject is psychologically disturbed or telling a lie, we assume there are changes in exhalation muscle tension of vocal cords and resonance characteristics from the vocal tract to the lips, including the mouth and nasal cavity. Although these characteristics are not sufficiently investigated yet, it is hypothesized that a guilty person's utterance to a critical question is different from his utterance to a control.

Based upon this assumption, detection deception through analysis of voices in respect to showing the pitch, intensity, duration and the sonagram has been studied and reported here.

Method

Subjects

A pilot study with a mock crime did not produce enough stress during the examination; therefore, materials taken from criminal cases were used for analysis. These materials cover 3 subjects who were confirmed as criminals by confession or by medical jurisprudence examination. The crimes involved were larceny (pick pocket, intrusion) and rape. The subjects were males ranging in age from 24 to 30.

Recorder

A Sony ECM-21 microphone was placed approximately 30cm from the subject's mouths. The voice was transmitted through a unidirectional condenser microphone to a Sony TC-777A tape recorder in the next room. The recording was made at a speed of 19cm/second. The recording sensitivity was adjusted by using a UV meter to monitor subject's voice during the pre-test interview. This sensitivity of each subject was maintained through the examination.

Procedures

The subjects were taken into a semi-sound proof room and were given a pre-test interview. POT and comparison question tests (CQT)[1] were administered to the subjects in accordance with the standard procedure using a Takei TRP-L polygraph. During the examinations, verbal responses of the subjects were recorded in the next room by a tape recorder.

Instrument processing

When the answers to CQT and POT questions were inconsistent, they were not analyzed; for example, when one answered "no" to a certain question in a series and later said "I do not know." As a result, we selected 21 questions from the 7 charts on subject no. 1, 27 questions from the 11 charts of subject no. 2, and 27 questions from the 11 charts of subject no. 3. In all, there were a total of 75 answers to be analyzed for voice pitch, intensity and duration.

The recorded voices were reproduced by Toshiba GT-710 tape recorder and the output directly connected to a Nippon Electronics PI-3A pitch intensity indicator. The reproduced level was - □ ~ Cbd which is the sound pressure measuring range of the indicator. Therefore, it was adjusted so that the maximum volume of the subject's voice was about -5db. The reproduction level of each subject's voice was kept constant throughout the analysis time. The pitch indicator was set to measure the changes in the range of 9Q - 360 Hertz. The output of pitch intensity indicator was recorded at 100 mm/s on a sheet of Yokokawa EMC-61 electromagnetic oscillograph.

The verbal responses of subject No. 1 were used exclusively in the sonagraph analysis. The materials were limited to those showing conspicuous deception reaction to the relevant questions on the polygraph charts. Finally, voice responses to 15 series of the 5 question lists were used. Voices were reproduced by an Akai 910 tape recorder and put into Kay's sonagraph 662B. The analysis band was set at 0 - 6 kHz and the analysis filter was set at 300 Hz.

Analysis materials

(l) A duration was calculated for each answer based on records of the pitch indicator. That is, the time from 0 Hz before the subject answered and back to 0 Hz after the answer was measured.

(2) A pitch was first analyzed by using the records obtained by the pitch indicator, but the difficulty was in determining what the changes of characteristics in the pitch were products of deception. Therefore, the highest point of the pitch of the subject's answer was picked up and its frequency was measured. The highest point of each answer in each series usually appeared in the same location. For example, in the answer wakarimasen (I do not know) of each series, the highest point of pitch was recorded at "ri" of the answer "wakarimasen", except in an unusual case. When the highest point reached was at "se" in some cases, the measurement was taken at that point.

(3) The intensity was analyzed as in the case of pitch analysis, but because of the unknown criterion for judgement and non-linear recording of intensity on the paper, the analysis was very difficult. The record paper showed 5mm difference between -40db and -35db, but showed 15mm difference between -15db and –l0db. In the next method, a graph

[1] * The Comparison Question Test is R/I, rather than a Control Question Test. Hence the initials CQT are not used in this text as they are commonly used in the United States and Canada. [Ed.]

Podlesny, J. A., & Raskin, D. C. Physiological measures and the detection of deception. Psychological Bulletin, 1977, 84. 782-799.

Reeves, T. E. The measurement and treatment of stress through electronic analysis of subaudible voice stress patterns and rational-emotive therapy. Unpublished doctoral dissertation, Walden University, 1976.

Rice, B. The new truth machines. Psychology Today. June 1978, 12, 61-64 ff.

Roessler, R., & Lester, J. W. Voice predicts affect during psychotherapy. Journal of Nervous & Mental Disease. 1976, 163, 166-176.

Smith, O. A. The measurement of anxiety: A new method by voice analysis. IRCS (Research on: Biomedical Technology, Psychiatry, and Clinical Psychology). 1974, 2. 1707.

Smith, G. A. Voice analysis for the measurement of anxiety. British Journal of Medical Psychology, 1977, 50, 367-373.

Tursky, B., Schwartz, G. E., & Crider. A. Differential patterns of heart rate and skin resistance during a digit-transformation task. Journal of Experimental Psychology, 1970, 83. 451-457.

Waskow, I. E. The effects of drugs on speech: A review. Psychopharmacology Bulletin, 1966, 3, 1-20.

Wiegele. T. C. The psychophysiology of elite stress in five international crises: A comparative test of a voice measurement technique. International Studies Quarterly, 1978, 22, 467-511.

Wiggins, S. L., McCranie, M. L., & Bailey. P. Assessment of voice stress in children. Journal of Nervous & Mental Disease, 1975, 160, 402-408.

Williams, C. E., Stevens, K. N. Emotions and speech: Some acoustical correlates. Journal of the Acoustical Society of America, 1972, 52, 1238-1250.

Worth, L. W., Lewis, B., & Raborn, G. W. Presence of the dentist: A stress evoking cue? Virginia Dental Journal, 1973, 52, 23-27.

Zajonc, R. Social facilitation. Science, 1965, 149, 269-274.

Reference Notes

1. Dektor Counterintelligence and Security, Inc. Psychological Stress-Evaluator (sales brochure), 5508 Port Royal Fd.. Springfield, VA 22151; Bell, A. D., Jr., Ford, W. H., ft -McQuiston, R. Physiological response analysis method and apparatus. United Stales Patent 3,971,034, July. 1976.

2. California Penal Code, Section 637.3, Chapter 1251 (Enacted September. 1978).

3. Kubis; J. F. Comparison of voice analysis and polygraph as lie detection procedures (Tech. Rep. LWL-CR-03B70). Final report. Contract DAAD05-72-C-Q2I7, U.S, Army Land Warfare Laboratory, Aberdeen Proving Ground, MD 21005.

4. McGlone. R. E. Tests of the Psychological Stress Evaluator (PSE) as a he and stress detector. Proceedings, 1975 Carnahan Conference on Crime Countermeasures, Lexington. K-Y. 1975.

5. Worth, J. W., & Lewis, B. J. An early validation study with the Psychological Stress Evaluator (PSE). Unpublished manuscript, Washington & Lee University, Lexington, VA. 1973.

6. Brenner, M. Stagefright and Strven's Law. Paper presented at the meeting of the Eastern Psychological Association, April 1974.

7. Older, H. J., & Jenney. L. L. Psychological stress measurement through voice output analysis. Planar Corporation, 4900 Leesburg Pike, Alexandria, VA 22303, 1975 (Final report, NASA Contract NAS 9-14146).

8. Rockwell, D. A., Hodgson, M., & Cook, D. Psychological Stress Evaluator: An attempt at validation. Paper presented at the meeting of the Society of Biological Psychiatry, June 1976.

9. Heisse, J. W., Jr. Audio stress analysis: A validation and reliability study of the Psychological Stress Evaluator (PSE). Proceedings, 1976 Carnahan Conference on Crime Countermeasures, Lexington. KY, 1976.

10. McGlone, R. E..ft Holhen, H. Partialanalysisof acoustic signal of stressed and unstressed speech. Proceedings, 1976 Carnahan Conference on Crime Countermeasures, Lexington, KY, 1976.

11. Brenner, M.. & Branscomb, H. H. Psychological Stress Evaluator Technical limitations affecting lie detection. Testimony presented at hearings on Senate Bill 1845 (to prohibit lie detection in employment settings). United Stales Senate, Committee on the Judiciary, Subcommittee on the Constitution, Washington, D. C., September, 1978.

and 2) that the present instrument is subject to serious practical problems which raise doubts about the appropriateness of its use within lie detection.

Because the Experiment 2 results were so clearly positive, it is possible that the PSE measure simply does not respond well to low stress tasks such as employed in Experiment 1 and, perhaps, in other studies reporting negative findings. Technical problems such as low scoring reliability would contribute to this poor response by decreasing the available signal-to-noise response ratio for the measure. The Experiment 2 results suggest that the PSE analysis encompasses a potential vocal measure which, because it applies well across subjects, may reflect some basic property of the vocal system. These results suggest, also, that the measure may have an important property of being relatively inaccessible to unaided observation. It would be valuable to know more about the similarities between the analysis employed by the PSE and the similar pitch-perturbation analysis proposed by Lieberman (1961). Kuroda, Fujiwara, Okamura, and Utsuki (1976), in addition, have employed a form of pitch-perturbation analysis to indicate differential stress reactions present in the radio-transmitted statements of pilots involved in critical aviation situations. These three similar approaches—by Lieberman, Kuroda et al., and the PSE—may converge on a breakthrough in vocal stress analysis, a breakthrough which may permit for the first time a practical psychophysiological measure based on the voice.

At the same time, it seems impossible to employ the present instrument while ignoring the real-life lie detection applications for which it is being sold. Detection of deception is an extremely complex form of stress analysis, which requires careful analysis of individual responses and requires a critical determination that the observed responses are caused by deception and not by other forms of psychological stress. In the case of the PSE, the single problem of scoring reliability is sufficiently serious to raise questions about any specific lie detection decisions. The problem of response-word noted in Experiment 2, and the possibility of conscious control suggested by Experiment 1, suggest further difficulties for practical examinations (Brenner & Branscomb, Note 11). The latter possibility, which seems plausible given the strong conscious control of the voice, should certainly be further tested (perhaps using biofeedback techniques). It is not surprising that technical limitations exist in a device as new as the PSE. Problems due to recording quality and response-word, especially, are common for acoustical measures. What does seem surprising is the complete failure of the manufacturer to note these problems in its uncritical efforts to sell the device for lie detection applications. Hopefully, the present data will stimulate interest among psychophysiologists to consider vocal measures in research on emotion and stress and thereby help to uncover and develop the valid underlying parameters apparently tapped by devices such as the PSE.

Acknowledgements

Experimental work was supported by a grant from the Roche Psychiatric Service Institute to Gary E. Schwartz to examine psychophysiological measures of human relaxation and stress. We thank Baruch Fischhoff and Lewis R. Goldberg for comments. Address requests for reprints to: Malcolm Brenner, Psychology Department, University of Oregon, Eugene, Oregon 97403.

References

Barland, G. H. Use of voice changes in the detection of deception. Journal of the Acoustical Society of America, 1974, 55, 423. (Abstract)

Barland, G. H. Detection of deception in criminal suspects: A field study. Unpublished doctoral dissertation. University of Utah, 1975.

Borgen, L. A., & Goodman, L. 1. Voice print analysis of anxiolytic drug effects: Preliminary results. Clinical Pharmacology & Therapeutics, 1976, 19, 104. (Abstract)

Brockway, B. F. Situational stress and temporal changes in self-report and vocal measurement. Nursing Research, 1979, 28, 19-24.

Brockway, B. F., Plummer, O. B., & Lowe, B. M. Effect of nursing reassurance on patient vocal stress levels. Nursing Research. 1976, 25, 440-446.

Chapman, A. J. An electromyographic study of social facilitation: A test of the "mere presence" hypothesis. British Journal of Psychology, 1974, 65, 123-128.

Ellis, J. G., Ellis, J, L., & Reeves. T. R. Anxiety measurements in patients on dialysis: No difference between home or hospital. Carle Selected Papers, 1977, 30, 100-104.

Hecker, M. H. L., Stevens, K. N., von Bismarck, G., & Williams, C. E. Manifestations of task-induced stress in the acoustical speech signal. Journal of the Acoustical Society of America, 1968, 44. 993-1001.

Holden, C. Lie detectors: PSE gains audience despite critics' doubts. Science. 1975, 190. 359-362.

Horvath, F. An experimental, comparison of the Psychological Stress Evaluator and the galvanic skin response in detection of deception. Journal of Applied Psychology, 1978, 63, 338-344.

Inbar, G. F., & Eden, G. Psychological Stress Evaluators: EMG correlation with voice tremor. Biological Cybernetics, 1976. 24. 165-167.

Kahneman, D., Tursky, B.. Shapiro, D., & Crider, A. Pupillary, head rate, and skin resistance changes during a mental task. Journal of Experimental Psychology, 1969, 79, 164-167.

Kuroda, I.. Fujiwara, O., Okamura, N., & Utiuki, N. Method for determining pilot stress through analysis of voice communication. Aviation, Space, and Environmental Medicine, 1976, 47, 528-533.

Lieberman, P. Perturbations in vocal pitch. Journal of the Acoustical Society of America, 1961, 33, 597-603.

Lieberman, P., & Michaels, S. B. Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. Journal of the Acoustical Society of America. 1962, 34. 922-927.

Lippold, O. Physiological tremor. Scientific American. 1971, 224. 65-73.

Lykken, D. T. The validity of the guilty knowledge technique: The effects of faking. Journal of Applied Psychology, 1960, 44, 258-262.

Lykken, D. T. Psychology and the lie detector industry. American Psychologist, 1974, 29. 725-739.
New York Times. January 6. 1977, p. 18.

Orne, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. American Psychologist. 1962.17, 776-783.

O'Toole. P. The assassination tapes. New York: Penthouse Press, 1975.

least 2 error-free trials, up to 5 error-free trials, had been performed for each arithmetic operation (median number of trials = 24).[2] The intertrial interval was 30 sec and each arithmetic operation appeared twice per 8 trials (random ordering).

Data Reduction

PSE analysis was carried out only on error-free trials. The recording and transcription procedures were identical to Experiment 1, and the Experiment 1 judge scored the data. Recording quality was excellent, permitting a 6-point scoring of stress (0-5) for each pattern. For display purposes; all scores were multiplied by 20 to produce an arbitrary 0-100 summary scale.

Results

Fig. 1A summarizes the PSE results. The graphed bar for each addition operation averages 230-320 spoken responses (digits), and the bar for the Rp operations averages 1034 responses. The results of Experiment 2, unlike the results of Experiment 1, are clearly positive. PSE scores varied significantly across operation instructions (F(3/105)=7.3). They showed a graded increase which parallels previous psychophysiological results (linear trend weighted by operation magnitude, f(l/105)=16.6). PSE scores did not vary significantly across the corresponding baseline responses (F(3/105)=0.3), and in Fig. 1A these responses are collapsed into a single bar. The raw baseline scores were: 50.5, 48.4, 49.6, and 49.0 respectively for +0, +1, +3, and +4. It may be noted that there was a small nonsignificant drop from the baseline bar to the +0 bar (F((l/105)=2.2), which could reflect relaxation after the baseline levels of anticipation for an unknown task which might be expected within the present procedure.[3]

Figs. IB and 1C summarize data from two dependent measures, observed errors (1B) and self-report scores of nervousness (1C), which were employed as manipulation checks.

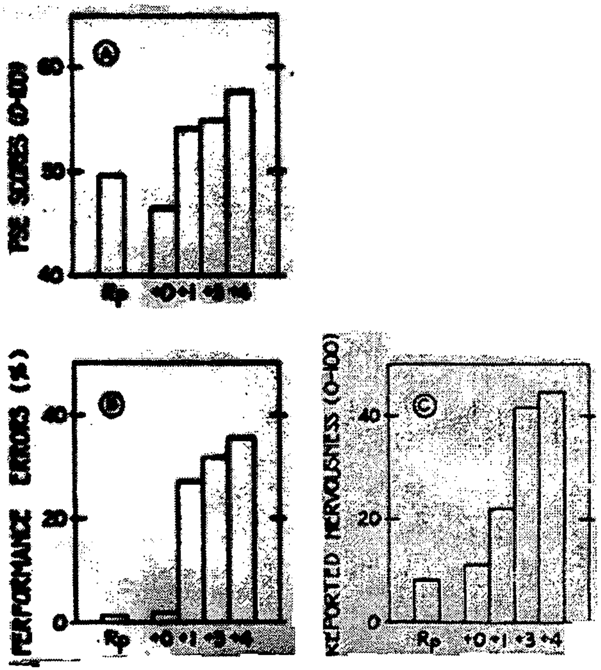


Fig. 1. Responses on the mental arithmetic task for three dependent measures: A) PSE scores of vocal stress, B) observed errors, and C) nervousness reported by subjects following the experiment. PSE data (A) include only trials on which subjects performed correctly.

The latter measure was derived from rating scales administered immediately after the experimental session (5 1/2 in. scales, anchors: "not at all stressed," "extremely stressed"). Both measures showed the predicted graded increase in response which also characterized the PSE scores.

Individual difference results in the PSE scores are summarized in Table 2. Subjects 12 and 6, the only subjects who performed the experiment without errors, showed lower baseline and operation scores than most of the remaining subjects (+4: +(14)=2.4).[4] More impressively, 15 out of 16 subjects showed a positive linear trend in response to the magnitude of the required arithmetic operation. Thirteen subjects showed their lowest operation score on +0(binominal-distribution p <.001). This distribution

[2] Five additional subjects failed to meet the criterion and were dismissed after 40 trials.

[3] An informal analysis was carried oul on 37 error (rials/chosen representatively across subjects and treatments. The average PSE score was 61.9 (cf. Fig. IA).

across subjects is completely unexpected for a vocal measure. In contrast, Hecker et al. (1968) report substantial individual differences on a battery of voice measures derived from spec-trographic analysis (voice amplitude, fundamental frequency, detailed waveform patterns of glottal pulses) when employed on a similar mental arithmetic task:"the manifestations of stress varied considerably from subject to subject." This single finding of applicability across subjects, especially given the known shortcomings of the PSE, provides the most compelling evidence from the experiment that some aspect of the PSE analysis is valid."

The Experiment 2 results indicated a serious artifact due to the linguistic structure of individual response words, with some response words showing characteristically high PSE scores and others characteristically low scores. The order was: "5" (average PSE score = 70), "9"(64), "4"(57), "1" (53), "0"(52), "7"(48), "3"(41), "2"(38).and "8" (35). This response-word effect was robust in -both baseline and operation responses, as robust, in fact, as the experimental manipulation (F(8/ 20)=9.6, 9.3 respectively). In the present design this effect was randomized across treatments, and, fortunately, it appears to have had little effect on the experimental results (each digit, inspected individually, showed a linear trend in response to the experimental manipulation). Such an artifact, however, would severely bias any PSE application using unrestricted responses.

TABLE 2

PSE summary scores for individual subjects on the mental arithmetic task

Subjects[a]	Rp[b]	Arithmetic-Operation Scores			
		+0	+1	+*3*	+4
2	62.8	58.9	61.7	72.5	77.5
3	60.1	48.0	66.7	58.3	60.0
5	59.1	57.9	62.0	72.0	66.3
8	56.6	44.0	57.8	65.3	56.0
7	54.S	58.0	55.0	53.3	60.0
1	52.0	50.0	51.3	54.7	53.3
4	30.7	46.3	64.0	50.0	72.5
10	50.7	44.3	48.3	49.5	47.5
16	50.1	43.2	58.7	58.3	45.0
14	49.8	47.0	58.9	49.0	67.5
12	48.4	49.0	53.0	41.3	45.0
9	48.3	52.6	57.0	61.3	61.0
6	42. 1	37.5	46.3	48.0	43.8
13	38.5	44.0	45.0	51.7	59.0
11	34.8	27.6	40.0	57.5	55.0
15	31.3	36.8	46.0	34.7	60.0

[a] Subjects ordered by the magnitude of baseline responses (Rp).

[b] Rp scores average 60-76 spoken digits.

[c] Arithmetic-operation scores average 8-20 spoken digits.

Finally, a listening test was carried out on Experiment 2 data which employed a sub-sample of 96 responses consisting of three +0 responses and three +4 responses from each subject. On a scoring of the PSE output records, three paid judges successfully differentiated +0 and +4 responses (t_1,=4.8, t_2=4.2, t_3=6.0 respectively, *df*=47)(this scoring followed a short training session). In a direct auditory scoring of the corresponding tapes, however, the judges failed to differentiate +0 and +4 responses significantly (t_1, = 1.3, t_2=0.1, t_3= -1.8; *df*=47). The judges showed a marginal agreement, however, in detecting which voice samples sounded stressful (r=.40, median value). Smith (1974), who employs a listening test similar to the present one, reports a marginal recognition of PSE-distinguished differences (65% recognition, chance =50%), while Lieberman and Michaels (1962) suggest that FM perturbations form a recognizable and important component of "natural sounding" speech. On the present test, in contrast, the changes discriminated, by the PSE were not readily recognized, or at least, were not readily coded as symptoms of stress.

Discussion

Experiment 1 and Experiment 2 provide opposing but complementary conclusions regarding the validity of the-PSE, and together effectively mirror the present validation literature. The available evidence suggests two conclusions: 1) that some aspects of the PSE analysis of stress are valid, suggesting the need for further studies employing parametric, multi-measure designs,

Method

Subjects

Fifteen male and 5 female students at Harvard University served as subjects. Each subject was paid $1.25.

Procedure

Subjects completed a questionnaire consisting of 10 items of personal information (e.g. "What is your mother's first name?" "What is the name of your first girlfriend/boyfriend?"), "and subsequently underwent a PSE interrogation based on the questionnaire answers. Following Lykken (I960), each subject was offered a reward to successfully conceal the correct answers from the interrogator ($5), and each was provided extensive information on the PSE and the interrogation procedure. Thus, for example, subjects were advised that it might be better to produce emotional responses to incorrect items rather than attempt to suppress actual emotional responses to correct ones.

For interrogation, the subject was seated in an IAC sound chamber (6 ft x 6 $^{1/2}$ ft), wearing earphones am blindfolded. A male interrogator, unaware of the subject's questionnaire answers, slowly read out loud each of the 10 questions plus 6 prospective answers to each. The subject repeated aloud all answers following the interrogator. In every case, the correct answer (i.e. the one supplied by the subject) was located randomly in positions 2 to 6. All remaining answers had previously been reviewed by the subject to eliminate any with emotional significance.

Data Reduction

All responses were recorded on a UHER recorder (4000-IC, 7$^{1/2}$ ips)by means of a B&K microphone. For each trial, an assistant transcribed the subject's five critical responses through the PSE-I (Mode III; 1 $^{7/8}$ or 15/16 tps). The resulting PSE records, identified by arbitrary code labels, were returned to the interrogator and each group of responses was rank-ordered for stress. There was missing data for the 4% of the responses which provided PSE records which could not be scored.

TABLE I

Rank-order distribution of guilty knowledge items as determined by PSE scoring

Rank-Order Categories	Proportion of Guilty Knowledge Items			
	Distribution Expected by Chance	Total Sample (167 trials)	6 Best Subjects (59 trials)	Best Trials (33 trials)
Highest observed stress	20%	19.8%	23.7%	15.2%
Second highest	20%	20.4%	27.1%	21.2%
Third highest	20%	21.0%	18.6%	21.2%
Fourth highest	20%	18.6%	15.3%	24.2%
Lowest observed stress	20%	20.4%	15.3%	18.2%
χ^2 (df = 4)		0.3	3.6	0.8

Results

Table 1 summarizes the group PSE results, and includes only trials for which full data were available. By random scoring the interrogator would classify an average 20% of the guilty knowledge items in each rank-order category. As shown in Table 1, the total results were close to chance levels: the interrogator classified only 19.8% of the guilty knowledge items in the highest stress category, 20.4% in the second highest category, and so forth. These results changed only slightly when the data were reduced to 6 subjects chosen by the interrogator for showing the clearest PSE responses. The results also failed to change

[1] Unless otherwise noted, all statistical tests employed a p< .05 rejection region.

when the data were reduced to only those trials ('best trials") for which the interrogator found a wide range in PSE response and most easy classification based on the principles of the PSE scoring. Although PSE patterns showed large variation, this variation was almost completely unrelated to the experimental manipulation.

Lykken (1960) reports a scoring procedure for individual subjects, based on deviations from a rectilinear distribution in a display such as employed in Table 1. These Lykken scores were computed for 8 subjects on whom complete data was available. The scores ranged from 1-4 with a median of 1.5, out of a possible range of 0-8. The distribution of scores was lower, although not significantly lower, than the distribution which would be expected from chance (Kolmogorov-Smirnov test: D = ,334, n = 8). An analysis of individual results, then, does not improve evidence for the PSE performance.

Discussion

The Experiment 1 results suggest a new failure for use of the PSE within interrogation settings (Notes 3,4; Barland, 1974,1975; Horvath, 1978). In part, this poor showing may reflect differences between the present procedure and the original procedure of Lykken (1960) (especially, a shortened list length). At the same time, there is a striking contrast between the present results and the results reported by Lykken. Using the GSR, he notes positive detection for 20 subjects out of 20 tested. The present results also contrast with those found by one of us (GES) who has successfully used the Lykken procedure for six years as a laboratory demonstration in his psychophysiology concepts and methods course. Experiment 1 argues that the PSE is not as sensitive a measure as the GSR (Horvath, 1978), at least in low-stress situations, and it argues that the PSE responses may be more vulnerable to conscious control.

Experiment 2

Experiment 2 employed the mental arithmetic task of Kahneman, Tursky, Subjects responded in time with clicks 9-12 and 29-32 respectively. There were 6 practice trials, and the experiment continued until at

Shapiro, and Crider (1969) and Tursky, Schwartz, and Crider (1970). In this task subjects perform arithmetic problems which vary in difficulty but must be executed under a fixed pacing schedule. Previous literature has shown graded increases in pupil dilation, heart rate, and GSR with increases in difficulty of the required problem.

Method

Subjects

Eleven male and 5 female Harvard students, paid $1.50 per half-hour, served as subjects.

Procedure

Each subject was seated in an IAC sound chamber, wearing earphones and facing a television computer terminal. A male experimenter sat quietly behind the subject throughout the experiment to maintain a moderate level of stress in the absence of physiological recording monitors (Chapman, 1974; Zajonc, 1965).

Stimulus presentation followed the procedure of Tursky et al). (1970) and was coordinated by a PDP-15 computer. Every trial began with a baseline problem ("Rp") in which the subject repeated out loud 4 digits presented serially on the terminal. The trial continued with a mental arithmetic problem: 4 new digits were presented serially, an operation was displayed, and the subject reported back the 4 digits adding either +4, +3, +1, or +0 to each digit. Subjects were instructed to truncate all answers to one digit (i.e. "12" became "2"). Stimuli were generated randomly from "0"to "9" with a stipulation .that "6" was never the correct response (in pretests "6" generated a PSE record of insufficient length to be scored). The pacing of the trial was time-locked by 32 audible clicks presented on the earphones: stimuli for the baseline problem appeared serially on clicks 5-8, stimuli for the arithmetic problem on clicks 17-20, and the operation instruction (e.g. "ADD / 1") on clicks 25-26.

Psychological Stress Evaluator- Two Tests of a Vocal Measure

Malcolm Brenner, Harvie H. Branscomb, and Gary E. Schwartz

Abstract

The Psychological Stress Evaluator (PSE), a commercial lie detector employing voice analysis, was tested on two laboratory tasks. On the guilty knowledge task of Lykken (1960), 20 subjects were interrogated on personal information after being offered a reward to fool the interrogator. PSE analysis failed to identify correct responses beyond chance levels. On the mental arithmetic task of Kahneman, Tursky, Shapiro, and Crider (1969) and Tursky, Schwartz, and Crider (1970), 16 subjects performed arithmetic problems which varied in difficulty but were performed under identical pacing. According to PSE scoring, stress increased with task difficulty. In addition, the PSE-measured differences occurred with high consistency across subjects. Some aspects of PSE analysis may be valid for the measurement of stress, although the validity of the analysis for practical lie detection is questionable.

Descriptors

Psychological Stress Evaluator (PSE), Voice analysis, Stress, Detection of deception, Unobtrusive measurement, Guilty knowledge paradigm, Mental arithmetic paradigm.

Attempts to develop a vocal measure of psychological stress, a measure which is truly unobtrusive, have proven disappointing. Although several parameters of the voice can be defined to respond to manipulations of stress, the relationship between these parameters and stress tends to vary widely across subjects and situations (Hecker, Stevens, von Bismarck, & Williams, 1968; Waskow, 1966; Williams & Stevens, 1972; Roessler & Lester, 1976; Podlesny & Raskin, 1977).

The present paper reports the results of two experiments using a new and controversial vocal stress measure, the Psychological Stress Evaluator (PSE), which is the original and most widespread of the recent vocal lie detectors (Holden, 1975; Podlesny & Raskin, 1977; Rice, 1978). According to the manufacturer (Dektor CI/S, Springfield, Va.), the PSE responds to an inaudible 8-14 Hz frequency modulation (FM) in the vocal signal, whose magnitude correlates inversely with stress and results from fine microtremors in the vocal muscles (Note 1). Voice samples are recorded on a UHER 4000-IC tape recorder and played at reduced speed through the PSE. The machine plots on heat-sensitive paper a filtered, time-based record of each speech utterance. This PSE record must be scored subjectively.

The PSE is sold as a lie detector and is widely used in employment screening. Much of the device's notoriety results from its potential use for unobtrusive measurement. The manufacturer, for example, includes a telephone hookup for the equipment which can be used in covert testing (the State of California, in pioneering legislation (Note 2), has prohibited vocal lie detection without prior written consent from the person being tested). Sensational press reports surround the PSE, including an attempted posthumous lie detection on Lee Harvey Oswald (O'Toole, 1975) and, incredibly, an abortive attempt by a Congressional subcommittee to employ secret voice testing on its witnesses (New York Times, 1977).

Validation evidence on the PSE is mixed. Ironically, some of the poorest results are reported in studies employing lie detection tasks. Failures of the PSE to respond beyond chance levels are reported by Kubis (Note 3) for a simulated robbery situation, and by Barland (1974), McGlone (Note 4), and Horvath (1978) for laboratory number-choosing and card-choosing tasks. In the

Horvath study, simultaneous GSR measurement permitted significant discrimination not provided by the PSE. Worth and Lewis (Note 5), in a study sometimes, cited as support for the PSE, report first-place correct calls ranging from only 58% to 8% (chance = 25%) on a card-choosing task. The most promising lie detection finding is reported by Barland (1974), who notes a positive relationship between PSE scores and polygraph scores on real-life interrogations. Positive discrimination by the PSE did not appear, however, in a more extensive field experiment (Barland, 1975).

Validation evidence is generally more favorable to the PSE for experiments which employ tasks other than lie detection (Smith, 1974, 1977; Brenner, Note 6; Wiggins, McCranie, & Bailey, 1975; Worth. Lewis, & Raborn, 1975; Older & Jenney, Note 7; Borgen & Goodman, 1976; Brockway, Plummer, & Lowe, 1976; Rockwell, Hodgson, & Cook, Note 8; Reeves, 1976; Ellis, Ellis, & Reeves, 1977; Weigele, 1978, Brockway, 1979). For example, Borgen and Goodman (1976), of the Parke-Davis Laboratories, report systematic changes in PSE scores to the Stroop color/word conflict task. These changes paralleled responses on a battery of psychophysiological measures. Rockwell, Hodgson, and Cooke (Note 8), in a study employing the anti-anxiety drug Librium, note significant correlations between change scores derived from the PSE and those derived from questionnaire scales. Smith (1974) reports a significant decrease in PSE scores, and in the GSR, following a 10-min relaxation period. Unfavorable findings are also reported for the PSE measure, however. An example is provided by Older and Jenney (Note 7), in a study sponsored by NASA, who report negligible changes for PSE scores in the radio-transmitted voices of Skylab astronauts as a function of presumed workload.

One issue raised frequently in validation evidence is the presence of problems not identified by the manufacturer. The most serious of these is the high subjectivity of PSE scoring, reflected in a distribution of reported reliability coefficients which are surprisingly low and erratic. For tests of interjudge reliability, the reported coefficients are: r = .89 (Note 8), r = .55 (Note 5), r = .39 (Note 7), and r = .38 (Horvath, 1978). For tests of split-half reliability, the coefficients are similarly low: r = .82 (Smith, 1974) and r = .39 (Smith, 1977). Scoring difficulties appear to be at least as prevalent among experienced judges and PSE instructors as among regular judges (Heisse, Note 9), suggesting that unreliability is built into the basic scoring procedure and not simply a reflection of inexperience.

A second problem reported for the PSE is a potential artifact due to recording quality. This problem is noted by Older and Jenney (Note 7), who report that the average PSE scores increased directly with the quality of available recordings (classified "good," "fair," or "poor"). A final problem is suggested by McGlone and Hollien (Note 10) who seriously question the microtremor explanation (Lippold, 1971) adopted by the manufacturer. Two attempts to test this explanation, using direct EMG monitoring of the major vocal muscles, have provided opposite results (McGlone & Hollien, Note 10; Inbar & Eden, 1976). It should be noted, however, that a disallowance of the microtremor explanation would not necessarily discount the entire explanation offered by the manufacturer. A relationship between FM perturbation in the voice and psychological stress has been previously noted by Lieberman (1961), and this relationship might be caused by several physiological mechanisms other than microtremor.

Against this mixed background, then, the present research attempts to provide a clearer profile of the PSE by testing it on two standard psychophysiological tasks. It was felt that positive results on either task might justify more extensive psychophysiological research on the PSE in the measurement of human stress responding. All scoring was carried, out by a trained PSE investigator, and the experimental conditions were strictly controlled to guarantee that all PSE analysis was blind.

EXPERIMENT 1

Experiment 1 employed the guilty knowledge task of Lykken (1974), developed as an interrogation procedure more sensitive than traditional lie detection procedures. Lykken (1960) reports strong detection of guilty knowledge items by means of GSR responses, even though subjects were offered relevant information and a monetary reward to conceal their correct answers.

arousal is an unanswered question. If, as its inventors claim, the PSE has been effective in stress identification, it is probable that the strong placebo effect of such an instrument has been the chief factor behind any significant accuracy results.

A situation is needed which very clearly causes physiological arousal, and does not rely simply upon an individual's self report of arousal. Since polygraphic measure have been used as indicators of various physiological parameters (Grossman, 1967), it seems feasible to use them as criteria of physiological arousal. A future study might investigate the PSE in comparison with other physiological measures, to establish if it is dependent on some minimal level of stress in order to be effective.

Acknowledgments

Dektor Counterintelligence & Security, Inc., 5508 Port Royal Rd., Springfield, Virginia, U.S.A.

Requests for reprints should be addressed to Brian E. Lynch, Forensic Science, Royal Ottawa Hospital, 1145 Carling Ave., Ottawa, Ont., Canada K1Z 7K4.

Reprinted from the Canadian Journal of Behavioural Science, 1979, 11, 89-94, with the kind permission of the authors and publisher.

Reference Notes

1. Barland, G.H. Use of voice changes in the detection of deception. Paper presented at the meeting of the Acoustical Society of America, Los Angeles, October 1973.

2. Kradz, M.P. The Psychological Stress Evaluator: A study. Unpublished manuscript, 1972 (available from Dektor.) Personal communication, July 1977.

3. Borgen, L.A. & Goodman, L.I. Voice analysis of anxiolytic drug effects: Prelimintary results. Paper presented at the American Society of Clinical Pharmacology and Therapeutics. Seattle, March 1976.

4. Brenner, M. Stagefright and Steven's Law. Paper presented at the Eastern Psychological Association Convention. New York, April 1974.

5. Smith, G.A. Analysis of the voice: A study. Unpublished manuscript, 1973. (Available from Department of Psychology, Powick Hospital, Worcester, England).

6. Dektor Counterintelligence and Security, Inc. PSE orientation course, 17 March 1976.

7. Dektor Counterintelligence and Security, Inc.

References

Grossman, S.P. A textbook of physiological psychology. New York: John Wiley, 1967.

Kubis, J.F. Comparison of voice analysis and polygraph as lie detection procedures. *Polygraph,* 1974, 3, 1-48.

Lippold, O.C.J. Physiological tremor. *Scientific American,* 1971, 224, 65-73.

Podlesny, J.A. & Raskin, D.C. Physiological measures and the detection of deception. *Psychological Bulletin,* 1977, 84, 782-799.

Reeves, T.E. The measurement and treatment of stress through electronic analysis of subaudible voice stress patterns and rational-emotive therapy. Unpublished doctoral dissertation, Walden University, 1976.

Stelmack, R.M. & Leckett, W.J. Effect of artificial pupil size on recognition threshold. Perceptual and Motor Skills, 1974, 39, 739-942.

Vetter, C. The lie machine. *Playboy*, 1973, April, 92.

Wiggins, S.L., McCranie, M.L., & Bailey, P. Assessment of voice stress in children. *The Journal of Nervous and Mental Disease*, 1975, 160, 402-408.

Worth, J.W., & Lewis, B. Presence of the dentist: A stress evoking cue? *Virginia Dental Journal*, 1975, April, 22.

Method

Subjects

The sample consisted of 43 university summer students ranging in age from 18 to 50, with a mean age of 26.1 years. There were 21 males and 22 females, representing a cross-section of socio-economic levels in a bilingual university environment. Because of the design utilized, all students constituted the experimental group without the necessity of a control group.

Apparatus

The stimuli consisted of 10 neutral words (at, by, cup, home, on, or, over, run, sky, the) and 10 taboo words (cock, cunt, fag, frig, fuck, prick, puke, screw, shit, tit; cf. Stelmack & Leckett, 1974), printed on a 7.5 x 12.5 cm cards with 20-pt Helvetica medium (capitalized) Letraset lettering. An additional neutral word (pen) was added as an initiating "damper" stimulus. Voice recording was taken on a Uher 4000 report I-C tape recorder using a Uher dynamic microphone M 136 and Scotch AV-177 low-noise tape. The tape recording was subsequently played into the Psychological Stress Evaluator (PSE-101) at speeds of either 4.7 cm/sec or 2/4 cm/sec, and filtered through Mode III.

Procedure

Before the experiment, all students completed the Eysenck Personality Inventory (EPI). Each student was then given a stack of 10 randomly arranged neutral and taboo word cards, plus the initiating neutral words. The random order was accomplished by blindly drawing each set of 10 cards from a box containing all 20 cards. Each student was asked to recite the words into the tape recorder after the experimenter had left the room. When finished, each student was asked to rate the 10 words on a 7-point rating scale, ranging from very pleasant to very disgusting. All recorded word lists were then processed on the PSE and distributed to 2 trained analysts and 10 untrained analysts for stress analysis. All raters used a rating chart composed of voice patterns identified by the Dektor Corporation (Note 6) as indicative of stress. None of the raters was aware of the type of words, or the proportion of neutral to taboo words. They were instructed only to compare the 430 word patterns and the rating charts to see if any of the patterns were similar.

Results

Table 1 presents the decisions made by each of the analysts on the 430 voice patterns, of which 216 were taboo words and 214 were neutral words. There were no statistically significant differences between the analysts on accuracy of rating (t(ll) = .62, p greater than .05). Both trained and untrained analysts were unable to discern differences in voice patterns between taboo and neutral words. That is to say, they were unable to sort the voice-stress patterns consistently, at a greater than chance level, into those that belonged with taboo words and those that belonged with neutral words.

In addition, there was no relationship between the analysts' pattern identifications and their resultant accuracies (r = -.01, biserial coefficient). Thus the total number of stress pattern identifications was not a predictor of accuracy outcome. The mean EPI results were within normal limits for university students (E = 11.2, N = 10.4, L = 3. 3). There were no significant correlations between word ratings and any of the EPI scales. There was a statistically significant difference between the student's rating of taboo words and neutral words (t(42) = 5.78, p less than .001).

TABLE 1

Breakdown of percentages in stress pattern identification

	Taboo words		Neutral words		Stress and neutral
Analyst	"Stress" (%) (True-Positive)	"No stress" (%) (False-Negative)	"Stress" (%) (False-Positive)	"No stress" (%) (True-Negative)	Correctly identified (%)
1. T[a]	41	59	51	49	45
2. T	67	33	78	22	45
3. UT[b]	15	85	15	85	50
4. UT	92	8	92	8	50
5. UT	64	36	57	43	54
6. UT	69	31	67	33	51
7. UT	75	25	86	14	44
8. UT	77	23	69	31	54
9. UT	84	16	87	13	49
10. UT	98	2	96	4	51
11. UT	87	13	88	12	50
12. UT	1	99	0	100	50
TOTAL	64	36	65	35	49 *t* .62 ns

Discussion

These results indicate that pattern identification of voice stress resulting from the utterance of taboo and neutral words was a chance occurrence. The analysts, regardless of training, performed at approximately chance levels in terms of accuracy of identification. Therefore, accuracy of pattern identification was not a function of extent of training in pattern identification. Since both trained and untrained analysts followed no consistent trend in identifying words, it must be concluded that pattern identification in this study was accomplished by random guessing. That is, the analysts were in no way consistent in their choice of patterns and, therefore, in their resultant accuracy.

The lack of significant difference between the actual accuracy rate and the expected accuracy rate may reflect, in part, a state of low level arousal when subjects uttered taboo words. Although the students rated the taboo words as significantly more disturbing than the neutral, the taboo words may still not have been sufficiently arousing to be picked up by the PSE. Since earlier studies have shown taboo words to be arousing, this explanation does not seem compelling. However, the inventors of the PSE (Note 7) suggest that it functions within limits of arousal which have not yet been defined. Thus, a certain level of arousal must be present in an individual in order for it to be picked up and displayed by the PSE. If this is the case, usage of such equipment in applied situations would require some external criterion measure of "sufficient arousal" before anything could be said about the voice pattern. With reference to the present study, if the uttered words were not registering on the PSE, then this would preclude any chance of correct identification by the stress analysts.

Many questions as to pattern identification, training effect, and minimum-maximum stress levels necessary with the PSE, are still unanswered. It is well known that the PSE is being used by police and private industry daily as a procedure for detecting deception. If, because of threshold activation limits, it cannot detect stress states equally on a continuum from no stress to maximum stress, then when and when not to use it without some other criterion measure of

References

[1] Rice, B.. "The New Truth Machines," Psychology Today. Vol. 12. No. I. June 1978. pp. 61-78.

[2] Brumlik, J. and Yap, C-B.. Normal Tremor: A Comparative Studv. Charles C Thomas, Springfield.Ill., 1970.

[3] Marshall, J., "Tremor." in Handbook of Clinical Neurologv. P.J. Vinken and G. W. Bruyn, Eds., Vol. 6, North Holland Publishing Co., Amsterdam. L968, pp. 809-825.

[4] Freund, H. J. and Dietz, V., "The Relationship Between Physiological and Pathological Tremor." in Physiological Tremor, Pathological Tremor and Clonus. Vol. 5 of Progress in Clinical Neurophysiology. J. E. Demsmedt, Ed., Karger, Basel, 1978. pp. 66-89.

[5] Horii, Y., "Some Statistical Characteristics of Voice Fundamental Frequency." *Journal of Speech and Hearing Research*. Vol. 18, No. 1, March 1975, pp. 192-201.

[6] Shipp, T., Fishman, B. V., Morrissey. P., and McGlone, R. E., "Method and Control of Laryngeal EMG Electrode Placement in Man," Journal of the Acoustical Societv of America, Vol. 48, No. 2. Pt.1. Aug. 1970, pp. 429-430.

[7] Vennard, W., Singing: The Mechanism and the Technique, Carl Fischer. Inc., New York, 1967.

[8] Seashore, C. E., Psychology of Music. Dover Publications, Inc., New York. 1967.

[9] Inbar, G. F. and Eden, G., "Psychological Stress Evaluators: EMG Correlation with Voice Tremor." *Biological Cybernetics*. Vol. 24, No. .3, Nov. 1976, pp.. 165-167.

Address requests for reprints or additional information to:
Thomas Shipp, Ph.D
Speech Research Laboratory (126)
Veterans Administration Medical Center
San Francisco, Calif. 94121

A Validity Study of the Psychological Stress Evaluator

Brian. E. Lynch and Donald R. Henry

Abstract

The Psychological Stress Evaluator (PSE) was assessed for its ability to display and detect arousal in the spoken word. Forty-three university summer students were asked to read aloud 10 words composed of random proportions of taboo and neutral words. PSE recordings of these words were then given to 2 trained and 10 untrained analysts for identification of stress patterns. Results indicated that, although the students rated the taboo words significantly more arousing than the neutral, the accuracy of identification of such words was no greater than chance for all analysts, regardless of training. It was concluded that the PSE may not be as effective as its manufacturers claim. Additional research appears warranted.

The Dektor Corporation of Springfield, Virginia has marketed an instrument called the Psychological Stress Bvaluator (PSE) which is claimed to measure stress, arousal, or physiological change associated with the voice, without the need of attached sensors. Traditionally, physiological measurement has used attached sensors with the result that a certain percentage of the measured arousal is artifically induced. If one is attempting to measure the degree of arousal or physiological change associated with a specific stimulus, then measurement without sensors would eliminate the possibility of sensor-induced arousal.

The PSE employs tape-recorded speech for the purpose of voice analysis, Briefly, the system involves feeding recorded vocalizations into the PSE to produce a visually observable medium. This medium or wave form is carefully analyzed in an attempt to identify frequency components of the recorded utterances that indicate physiological manifestations of psychological stress. More specifically, the PSE is intended to record the frequency components of uttered speech in such a way that purported infrasonic variations become indicators of the degree of stress. The Dektor Corporation suggests that these infrasonic variations are muscle microtremors occurring at 8-12 Hz (Lippold, 1971), and that the resultant patterns can be analyzed for stress using various modes (electronic filtering) and tape speeds.

PSE voice analysis has been researched in various ways. Barland (Note 1), Kradz (Note 2), Kubis (1974) and Vetter (1973) have used the PSE in the detection of deception, using mock and real crime situations. Borgen and Goodman (Note 3), Brenner (Note 4), Reeves (1976), Smith (Note 5), Wiggins, McCranie, and Bailey (1975), and Worth and Lewis (1975) have used the PSE in various experimental situations, ranging from psychotherapeutic effectiveness to stage fright. Podlesny and Raskin (1977) state that "at this point there appears to be no scientific evidence that PSE analysis yields accuracies as high as those obtained with standard polygraph procedures, and little evidence that results exceed chance levels" (p. 796).

Much of the research presently available on the PSE has lacked external truth criteria for validation requirement and also aid in the analysis. Emotionally powered words have been used in various physiological investigations as reliable laboratory inducers of mild stress (Stelmack & Leckett, 1974). The purpose of the present study was (a) to investigate the validity and inter-judge agreement of the PSE by assessing the rate of detection of arousal in spoken words; and (b) to see if naive analysts could analyze stress by matching to sample.

independent of pitch, with the oscillation occurring predominantly in the signal amplitude. In other words, the average rate of vocal oscillations in both singers and pathologic subjects, whether in frequency or amplitude, fell within the same value of 5 to 6 Hz. Moreover, in both groups the rate of oscillations was not affected by the pitch produced. Sustained phonation by the single non-singer subject showed random nonrhythmic variations in both voice frequency and amplitude.

Study 2

One young adult male volunteer underwent electromyographic (EMG) assessment of laryngeal muscle activity. The purpose was to sample electrical activity directly from critical muscles in the larynx to determine if there were periodic muscle contractions buried beneath the large electrical interference pattern picked up from these muscles as the subject produced phonation.

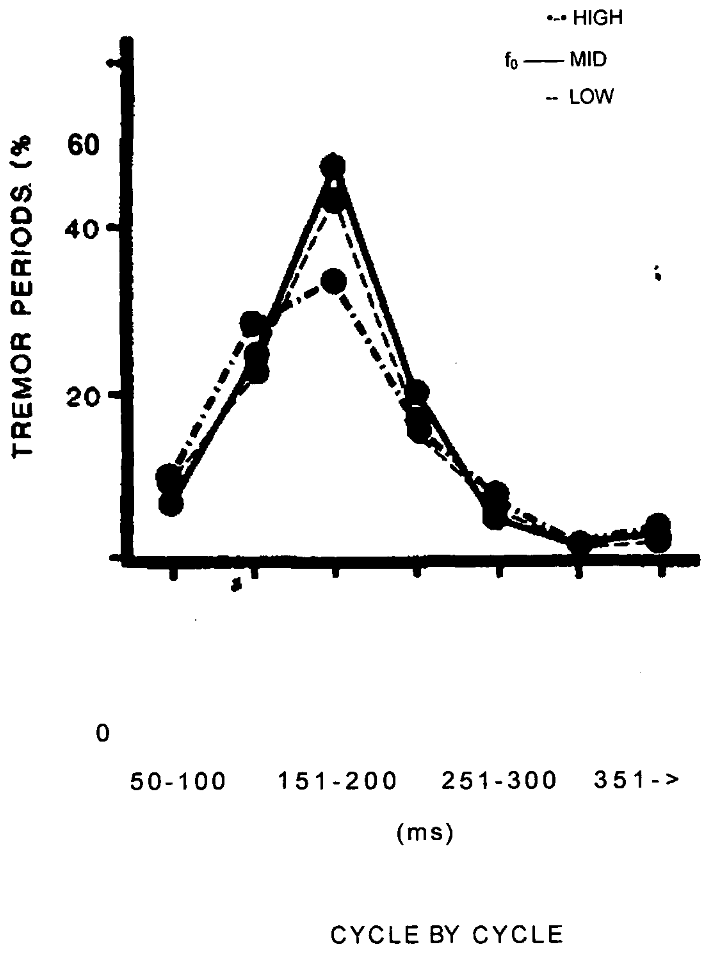


FIG. 2—*Percentage of each tremor cycle duration (in ms) at low, medium, and high fundamental frequency (fo) for 24 patients with vocal tremor during sustained phonation.*

Intramuscular hooked-wire electrodes were introduced to the target muscles [6]. Two hours after electrode insertion, EMG signals were recorded on frequency-modulated tape from the cricothyroid and the posterior cricoarytenoid muscles during conversational speech and during sustained phonation (isometric muscle contraction). Long before this time, the subject had adjusted to the experimental situation and produced voice and speech easily with no subjective or objective indications of stress. To verify the system's capability to discern normal tremor, EMG activity was also sampled from the biceps muscle while the subject maintained an isometric contraction with his forearm supinated at 90°.

The EMG recordings from larynx and limb muscles were subjected to fast Fourier analysis that revealed the spectrum of energy in the complex electrical signal. It was anticipated that normal physiologic tremor (microtremor), if present, would show up in the EMG analysis as an energy peak somewhere between 8 and 12 Hz. It was found that EMG activity during conversational speech changed so rapidly over time (to accommodate normal speech phonation patterns) that at the present sampling rate no Fourier analysis could be made of these signals. Analysis of a 1-s segment of the EMG activity from both the posterior cricoarytenoid and cricothyroid muscle during nonstressful sustained vowel phonation failed to reveal any periodic component in the frequency band from 1 to 20 Hz; the electrical energy was randomly distributed throughout the spectrum. A 1-s segment of the EMG activity from the biceps revealed a prominent energy peak at 9 Hz, indicating periodic contraction within the range of normal physiological tremor rate.

Discussion

The rate of vocal vibrato in singers and of vocal tremor in this study is consistent with values generated in other studies of these parameters [7,8]. It would appear that oscillatory contraction of laryngeal muscles, whatever the cause, averages about 6 Hz with considerable variability between 3.5 and 7 Hz. The finding of a periodic muscle contraction at around 9 Hz in a limb muscle is consistent with the neurophysiologic data on normal physiologic tremor and validates the instrumentation and analysis techniques used in this study. The failure to find a similar tremor-like muscle pattern in the laryngeal muscles is contrary to the study of Inbar and Eden [9], who reported tremors of 10 to 20Hz in muscle activity generated during sustained phonation. Their use of surface electrodes placed on the neck makes it difficult to be precise about the origin of the obtained EMG signals and, therefore, casts doubt on their conclusion that the obtained muscle patterns were, indeed, samples from a critical laryngeal muscle for phonation. Further evidence that their obtained signals were nonlaryngeal in origin lies in their correlations of obtained EMG with the frequency of the third formant, the location of which has little to do with laryngeal activity. Perhaps these investigators were presuming to sample some type of muscle activity that altered the shape of the supralaryngeal vocal tract instead.

Summary

These investigations demonstrate that laryngeal muscles can oscillate at rates between 4 and 7 Hz to produce frequency changes associated with vocal vibrato. Further, periodic muscle oscillations somewhere along the vocal tract in patients with vocal tremor produces a marked, rhythmic variation in amplitude of the voice signal at a rate from 3 to 7 Hz. The failure to find physiologic evidence of normal tremor in sampled laryngeal muscles casts some doubt on the assumption made by the manufacturers of stress analysis instruments that they are, indeed, detecting the presence of laryngeal muscle tremor.

Acknowledgments

The reported research was supported in part by National Institutes of Health Research Grants NS 14199 and 14729 and by the Veterans Administration Research and Development Service.

as having a fast rate of from 8 to 12 Hz, relatively small amplitude, and a continuous but rather irregular waveform. In contrast, abnormal tremor rate is from 3 to 8 Hz with variable amplitude and a waveform that can range from regular and rhythmic to irregular. Marshall [3] reported that tremor rate depends heavily on the anatomic site; tremor. amplitude is related to muscle load and tension [4] and, therefore, is a misleading quantifier of tremor. Muscles in midline structures such as the larynx are subject to possible oscillatory or tremor behavior, as are the more common limb muscles.

The nature of the nerve-muscle function in the larynx is so divided that alterations in nerve efferent impulses to one set of muscles will influence only vocal frequency while impulses to the others will principally influence the amplitude. This arrangement of muscle to function allows independent study of each acoustic parameter, for example, voice frequency changes during vocal vibrato in singers and voice amplitude changes in patients with vocal tremor.

The present study was designed to compare and contrast acoustic measures derived from the sustained vocalization of singers, patients with vocal tremor, and a normal nonsinger. In addition, electrical activity from limb and several laryngeal muscles during isometric contraction was also investigated in the normal subject. It was felt that such a combined investigation would define the characteristics of tremor when present in the voice and determine the presence or absence of small, rhythmic contractions in sampled limb and laryngeal muscles.

Study 1

The first portion of the study was concerned with specifying the rate, amplitude, and regularity of vocal vibrato in singers since this behavior can be considered caused by "normal" muscle contraction oscillations. The subjects were five men and five women who were members of an internationally known opera company. Each subject was tape-recorded while sustaining vocalization on the vowel |a| for 7 to 12 s at each of three pitch levels: low, medium, and high. At each pitch, vocalization was produced with high and low effort; thus, six vocalizations were analyzed for each subject for a total of 60 samples. The singers were instructed to produce their best quality vocalization for each trial. The exact frequency and relative effort level were not dictated; the singer selected a representative pitch and produced sustained phonation at that frequency using two distinctly different effort levels.

The pathologic subjects were 20 females and two males with a vocal disorder of unknown cause known as spastic dysphonia who, along with the primary symptom of a "strangled" voice, had an accompanying pronounced vocal tremor. These subjects were recorded sustaining the vowel |a| for as long as possible at low, medium, and high pitch within their range, but only at one comfortable effort level at each pitch.

The recordings for both groups of subjects and the normal nonsinger were played back through a graphic-level recorder, an instrument that makes sensitive measures of amplitude variations, and through a sound spectrograph displaying the acoustic spectrum with a special amplitude display. The recorded samples were also analyzed from an oscillographic display and by computer analysis [5] that provided a calculation of each pitch period from which any pattern of voice frequency change could be detected. Figure 1 shows typical graphic outputs produced by the analyzing instruments.

Results of Study 1

The results of Study 1 demonstrated that the principal difference in the acoustic output of the singer and vocal tremor groups was that frequency variation was responsible for vocal vibrato while amplitude was the primary variable in vocal tremor. As shown in Table 1. when all subjects and conditions were pooled for the singers the average vibrato rate was 5.4 Hz with a range from 4.7 to 6.6 Hz with little or no discernible amplitude fluctuation. Figure 2 shows that for pathologic subjects the tremor rate ranged from 3 to 10 Hz, with the dominant rate at 5 to 6 Hz (each tremor cycle = 151 to 200 ms)

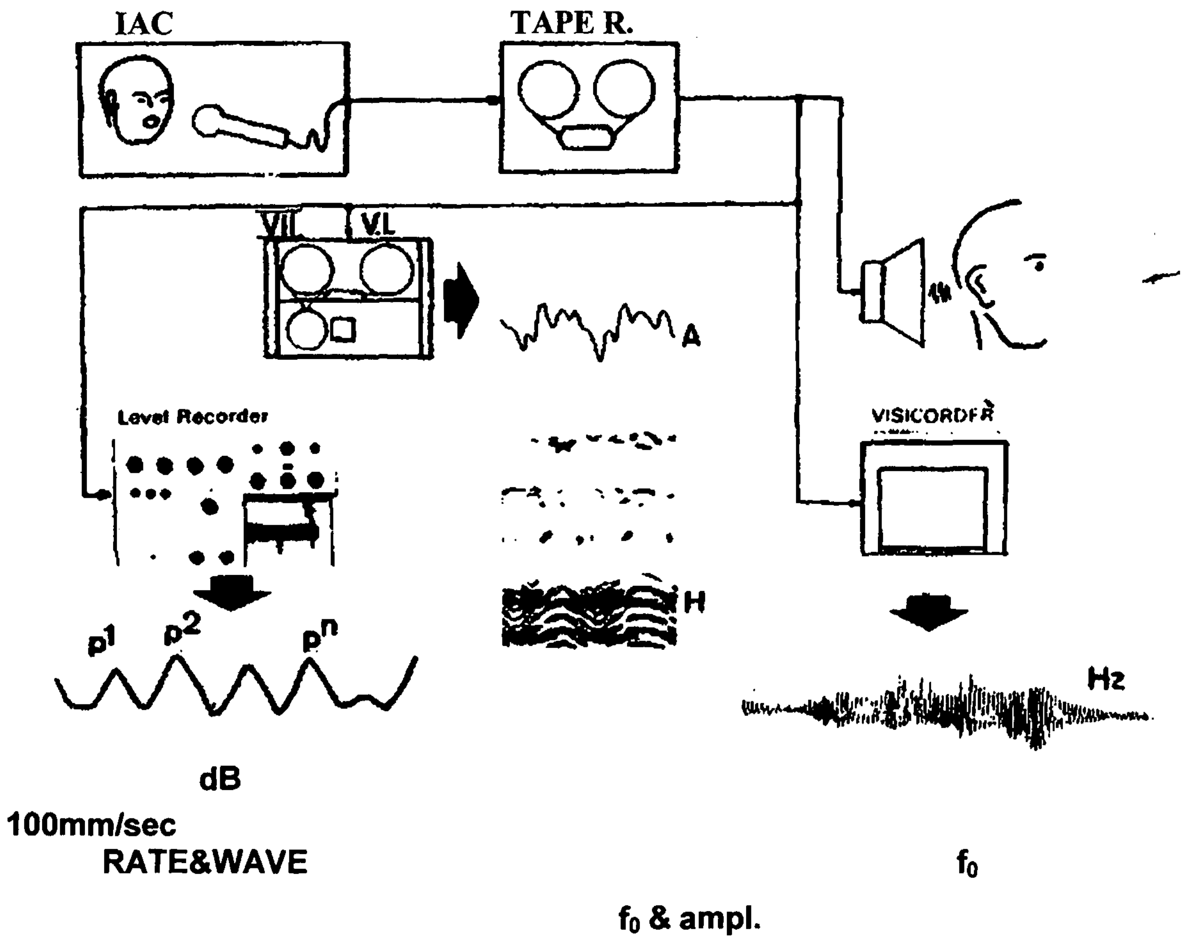


FIG. 1—*Instrumentation for recording and playing back subjects' voices to analyze vocal intensity (dB), amplitude (A), acoustic spectrum (H), and fundamental frequency (fo).*

TABLE 1—*Mean vibrato rate, in hertz, for five male and five female singers sustaining phonation at low, medium, and high pitch.*

Subjects	Low Pitch	Medium Pitch	High Pitch	Pooled
Males	5.5	5.4	5.4	5.4
Females	5.7	6.1	6.0	5.9

[26] Testimony of A. Bell and M. Kradz, Milano. F. v. Garrison et al., transcript of proceedings in the District Court of the United States for the Western District of North Carolina, Charlotte Division, C-C-79-334, Dec. 4 and 17, 1979, pp. 267.

[27] Report of tke Department of Commerce on the Feasibility and Desirability of Licensure of Audio Stress Examiners to tke Governor and the General Assembly of Virginia. House Document No. 5, Commonwealth of Virginia, Richmond. VA., 1981.

[28] Polygraph Control and Civil Liberties Protection Act, hearings before the Subcommittee on the Constitution of the Committee on the Judiciary, United States Senate, 95th Congress, 15-16 Nov. 1977 and 19 and 21 Sept. 1978, U.S. Government Printing Office. Washington. DC, 1978.

[29] Rice, B., "The New Truth Machines," *Psychology Today.* Vol. 12, No. 1, June 1978. pp. 61-78.

[30] Marston, W., The Lie Detector Test. Richard R. Smith, New York, 1938.

[31] Trovillo, P., "A History of Lie Detection," *Journal of Criminal Law and Criminology.* Vol. 29, 1939, pp. 848-881 and Vol. 30, 1939, pp. 104-119.

[32] Horvath, F., "Detection of Deception: A Review of Field and Laboratory Procedures and Research," *Polygraph,* Vol. 5, No. 2, June 1976, pp. 107-145.

[33] Podlesny, J. and Raskin, D., "Physiological Measures and the Detection of Deception," *Psychological Bulletin.* Vol. 84, No. 4, Dec. 1977, pp. 782-799.

[34] Horvath, F., "Polygraphy: Some Comments on the State of the Art," *Polygraph,* Vol. 9, No. 1, March 1980, pp. 34-41.

[35] Reid, J. and Inbau, F., Truth and Deception: The Polygraph ("Lie Detector") Technique, 2nd ed., Williams and Wilkins. Baltimore, 1977.

[36] Lykken, D., "Psychology and the Lie Detector Industry," *American Psychologist,* Vol. 29. No. 10. Oct. 1974, pp. 724-739.

[37] Ansley, N., Ed., "PSE User Loses in Federal Court," *American Polygraph Association Newsletter.* Vol. 14, No. 1, Jan.-Feb. 1981, p. 2.

[38] Ansley, N., Ed., "Illinois Supreme Court Upholds Polygraph Licensing Law," *American Polygraph Association Newsletter.* Vol. 13, No. 6, Nov.-Dec., 1980, p. 1.

Current Evidence for the Existence of Laryngeal Macrotremor and Microtremor

Thomas Shipp[1] and Krzysztof Izdebski[2]

Reference

Shipp, T. and Izdebski, K., "Current Evidence for the Existence of Laryngeal Macrotremor and Microtremor," Journal of Forensic Sciences. JFSCA,
Vol. 26. No. 3. July 1981, pp. 501-505.

Abstract

To test for the existence of laryngeal "microtremors" two experiments were conducted on humans. The first analyzed the acoustic characteristics of observable tremors (macro-tremors) in the voice of singers using vocal vibrato and in pathologic subjects producing vocal tremor. In both of these groups acoustic oscillations between 4 and 8 Hz were found. The second study, using a normal subject, sampled electromvographic (EMG) activity from laryngeal and arm muscles during isometric contraction to determine if a periodic component (microtremor) was present in either muscle's contraction pattern. A 9-Hz signal was detected in limb muscle contraction, whereas no periodicity was found in signals from laryngeal muscles. The application of these findings to the theory behind voice "stress" analyzers is discussed.

Keywords

pathology and biology, larynx, speech

The publicity and promotion surrounding the use of instruments designated as "stress" evaluators report that the presence or absence of laryngeal tremor in the voice is the basis for the determination of deception. Some of these instruments purport to perform this tremor analysis "on-line," presenting the results in a light display, while others record the voice signal and the resultant x-y chart printout is "analyzed" by someone using criteria set up by the manufacturer [*1*]. According to the developers' information, an unstressed voice has "normal" microtremors, while a stressed voice exhibits changes in the frequency of these rhythmic contractions. It seems implicit in the information provided by the manufacturers that these so-called normal laryngeal microtremors affect the voice output of the subject, and, though inaudible to the human ear, their presence or absence or a tremor frequency change can be detected by the "stress analysis" instrument. The physiologic basis for the existence of certain large types of oscillatory muscle behavior in the larynx is well known, while the presence of smaller tremor-like activity in this region is not well documented. This paper is intended to review the relevant evidence on laryngeal muscle tremor and to determine the reality of the acoustic and physiologic existence of such large and small tremors in the human larynx.

The phenomenon of rhythmic oscillations known as tremor is well described in both normal and pathologic subjects. The physical characteristics of limb tremor are the tremor's rate, amplitude, and overall waveform or pattern [*2*]. Normal tremor is described

Reprinted with permission from the Journal of Forensic Sciences, Vol. 27, No. 2, Copyright ASTM international, 100 Barr Harbor Drive, West Conshohocken, PA 19428.

Received for publication 21 July 1980; revised manuscript received 25 Oct. 1980; accepted for publication 29 Oct. 1980.

[1] Chief, Speech Research Laboratory, VA Medical Center, San Francisco, Calif., and associate professor, Department of Otolaryngology, University of California at San Francisco.

[2] Assistant professor, Department of Otolaryngology, and director, Voice Science Laboratory, University of California at San Francisco.

developmental history and maintain that voice stress analyzers represent advanced technology that, among other things, "simplifies chart-reading and greatly reduces both the training time required and the subjectivity of the chart reading [25, p. 64]." The evidence, of course, most clearly does not support such assertions. It is important to point out, however, that even if the evidence showed a dependable relationship between deception and what is recorded by voice stress analyzers, the historical, scientific, and practical lessons and developments in the lie detection field are proof enough of the falsity of such assertions as those made by proponents of voice stress devices. In other words, there can be no device, no instrument, no new technology that makes lie detection any less complex than it has already been shown to be.

In summary, the promise of voice stress analysis in the lie detection field is not and may never be a reality. All of the reliable evidence now available shows that none of the voice stress devices is useful in detecting deception; the fact that the precise relationship between the components of the voice spectrum and emotional states has not been adequately specified suggests a formidable obstacle to be overcome before analysis of the voice may prove of value in lie detection. The fact that voice stress devices have apparently been accepted rather uncritically by some law enforcement agencies, and for some forensic science purposes, is a development which, judging from the available evidence, cannot now be justified.

Acknowledgements

The opinions or assertions contained herein are the private views of the author and are not to be considered as official or as reflecting the views of the American Academy of Forensic Sciences or the Journal of Forensic Sciences. Address requests for reprints or additional information to Frank Horvath, Ph.D. School of Criminal Justice Michigan State University E. Lansing, MI 48824.

References

[1] Lippold, O., "Physiological Tremor," *Scientific American.* Vol. 224. No. 3, March 1971, pp. 65-73.

[2] *PSE Orientation Course Manual.* Dektor Counterintelligence and Security, Inc., Springfield, VA, undated.

[3] *Voice Stress Computer.* Omnitronics Research Corp., Akron, Ohio, 1980.

[4] Bennett, R., Hardee, P., and Klobert, R., *Hagoth: Fundamentals of Voice Stress Analysis*. Hagoth Corp., Issaquah, WA, 1977.

[5] *Voice Stress Analyzer System, Mark IX-P.* Communication Control Systems, Inc., New York, 1978.

[6] O'Toole, G., "Lee Harvey Oswald Was Innocent," *Penthouse.* Vol. 6, April 1975. pp. 45-46, 124-132.

[7] Jenkins, J., "Voice-Analyzing Lie Detector Yields Questionable Results," *Albuquerque Journal* (newspaper), 12 Oct 1979, p. F-2.

[8] Shipp, T. and McGlone, R.. "Physiologic Correlates of Acoustic Correlates of Psychologic Stress." presented at the meeting of the Acoustical Society of America, Los Angeles, Nov. 1973.

[9] McGlone, R. and Hollien, H.. "Partial Analysis of Acoustic Signal of Stressed and Unstressed Speech," in *Proceedings of the 1976 Carnahan Conference on Crime Countermeasures,* BU 110, ORES Publications, College of Engineering, University of Kentucky, Lexington, 1976, pp. 19-21.

[10] Inbar, G. and Eden, G., "Psychological Stress Evaluators: EMG Correlation with Voice Tremor," *Biological Cybernetics*, Vol. 24, 1976, pp. 165-167.

[11] VanDercar, D. H.. Greaner, J., Hibler, N. S., Spielberger, C. D., and Bloch, S., "A Description and Analysis of the Operation and Validity of the Psychological Stress Evaluator," *Journal of Forensic .Sciences*. Vol. 25, No. 1, Jan. I980, pp. 174-188.

[12] Brenner, M., Branscomb, H., and Schwartz, G., "Psychological Stress Evaluator: Two Tests of a Vocal Measure," *Psychophysiology* (Madison. WI). Vol. 16, No. 4, July 1979, pp. 351-357.

[13] Lynch, B. and Henry. D.. "A Validity Study of the Psychological Stress Evaluator," *Canadian Journal of Behavioral Science*. Vol. 11, No. 1, March 1979, pp. 89-94.

[14] Borgen, L. and Goodman, L., "Voice Print Analysis of Anxiolytic Drug Effects: Preliminary Results," *Clinical Pharmacology and Therapeutics*, Vol. 19, No. 1, March 1976, p. 104.

[15] Wiggins, C., McCranie, M.. and Bailey, P., "Assessment of Voice Stress in Children," *Journal of Nervous and Mental Disease*, Vol. 160, No. 4, Dec. 1975, pp. 402-408.

[16] Worth, L.. Lewis, B., and Raborn, G., "Presence of the Dentist: A Stress Evoking Cue?" *Virginia Dental Journal*, Vol. 52, No. 2, April 1975, pp. 23-27.

[17] Brockway, G., "Situational Stress and Temporal Changes in Self-Report and Vocal Measurement," *Nursing Research*. Vol. 28, No. 2, March 1979, pp. 19-24.

[18] Kubis, J., "Comparison of Voice Analysis and Polygraph as Lie Detection Procedures," report, Contract DAAD05-72-C-0217, U.S. Army Land Warfare Laboratory, Aberdeen Proving Ground, MD, Aug. 1973.

[19] Barland, G., "Use of Voice Changes in the Detection of Deception," presented at the 86th meeting of the Acoustical Society of America, Los Angeles, Oct. 1973 (erratum sheet issued March 1974).

[20] Barland, G., "Detection of Deception in Criminal Suspects: A Field Validation Study," Ph.D. dissertation, Dept. of Psychology, University of Utah, Salt Lake Ctty, 1975.

[21] Nachshon, I. and Feldman, B., "Vocal Indices of Psychological Stress Evaluator," *Journal of Police Science and Administration*, Vol. 8, No. 1, March 1980. pp. 40-53.

[22] Horvath, F., "An Experimental Comparison of the Psychological Stress Evaluator and the Galvanic Skin Response in Detection of Deception," *Journal of Applied Psychology*. Vol. 63, No. 3, June 1978. pp. 338-344.

[23] Horrath, F., "Effect of Different Motivational Instructions on Detection of Deception with the Psychological Stress Evaluator and the Galvanic Skin Response," *Journal of Applied Psychology*. Vol. 64, No. 3, June 1979, pp. 323-330.

[24] Horvath, F., "Detecting Deception in the Voice: The Validity of the Psychological Stress Evaluator," presented to the American Polygraph Association, St. Louis, Aug. 1978.

[25] Bell, A. "The PSE: A Decade of Controversy." *Security Management*, Vol. 25. No. 3, March 1981, pp. 63-73.

Because of the conflicts between Kradz's original report, what he has stated in testimony, and the recently distributed copy of his report, it is not possible to determine what Kradz actually did. The serious and unexplained methodological deficiencies in the Kradz study clearly indicate that that study does not meet generally accepted scientific standards; his reported findings, therefore, are of questionable value in assessing the validity of voice stress analysis.

One of the coinventors of the PSE has reponedly claimed that the device is 96.78 percent effective [7, p. F-2]." That claim apparently is based on a statistic reported by another voice stress proponent, Heisse, as a result of a study he carried out to investigate the "reliability and validity" of the PSE.[7] In his study, Heisse selected 53 cases (contributed by PSE users) in which the PSE was used to determine the truthfulness of the suspects (some of the "suspects" were applicants for employment, not persons involved in criminal investigations). Twenty-six of the suspects were apparently known to have been deceptive (to have shown "some form of deception") during their PSE testing; 27 were known to have been truthful. In each case ground truth apparently was established by a confession that indicated either the deception of the guilty suspect or the truthfulness of the innocent suspect. Of the 53 separate suspects tested, 25 of them were involved in three separate investigations.

Heisse asked 12 PSE users to evaluate the PSE charts of each of the 53 suspects and to determine whether each suspect was "truthful" or "deceptive." He reported his findings thusly: "There are 258 acceptable interevaluator replies. Among the replies there are 10 errors. . . . Hence, the interrevaluator reliability is 96.12 percent." Later in his paper he reports: "The compliance between evaluators and the known results with 258 evaluation replies is 96.12 percent. If examiners are included in this group . . . the reliability jumped to 96.78 percent."

Heisse's report, like that of Kradz, fails to disclose a number of important methodological details. Precisely how the cases were sampled, for example, is not revealed, nor is any procedure identified that would have ensured the independence of those persons who evaluated the response data. Since 25 of the suspects were involved in the same three investigations, it is certain that the tests carried out on those persons in each investigation were not independent; yet, Heisse does not indicate how that issue was dealt with, if at all. Although there are other serious methodological problems evident in the Heisse study, it is also the case that his findings were not correctly interpreted. Heisse himself, for instance, has stated that contrary to what his report suggests his findings only deal with the issue of reliability—how consistently his evaluators interpreted his data—and not with validity.[8] But, proponents often use his statistics to support their claim that voice stress analysis is 96.87% accurate. Judging from what was reported by Heisse and Kradz, such a claim is unfounded.

A final study claimed to support voice stirs analysis is a paper reported by Dahm,[9] who sent questionnaires to 423 users of the PSE; of those, 46 responded to questions about several characteristics of their use of the PSE. Dahm's major findings were reported as follows. First, he said that polygraph and PSE examinations were in agreement 5037 times in 5045 cases, "for a correlation of 99.84%." Second, "Based upon 10,202 PSE examinations . . . there was not one case in which the PSE had been found in error (28, p. 231]." It is, of course, clear that Dahm's data represent merely the unsubstantiated opinions of only a small number of PSE users; they are not sufficient to indicate whether or not voice stress analysis it a valid means of detecting deception.

[7] J. Heisse, "Audio Stress Analysis: A Validation and Reliability Study of the Psychological Stress Evaluator (PSE)," unpublished manuscript dated 1 Feb. 1976, available from the author, 144 Cliff St., Burlington, VT.

[8] J. Heisse, Burlington, VT. personal communication, 11 March I980.

[9] A. Dahm, "Study of the Field Use of the Psychological Stress Evaluator," unpublished paper distributed by Dektor, Inc., Springfield. VA., undated.

Thus, the Kradz, Heisse, and Dahm reports constitute at best merely testimonial, not scientific, evidence of the effectiveness of voice stress analysis. The merits of those studies notwithstanding, however, it is interesting that all of them were reported by proponents of voice stress analysis after 1971; neither the manufacturers nor the other proponents of voice stress devices have yet produced a report of research which was carried out before the devices were publicly marketed. The developmental research supporting the validity of the devices in lie detection is, curiously, not available. It is also important to point out that the findings in the proponents' studies regarding the accuracy of voice stress analysis have not yet been replicated in any objective, independent research. One manufacturer, asked for proof of the validity of his voice stress analyzer, reportedly sent to author B. Rice [29] a packet of ten studies, all of them very favorable. The studies were unpublished; two were apparently performed by an independent testing firm. When Rice investigated the firm he reportedly found that its president was the manufacturer of the voice analyzer. When that manufacturer was asked who did the other studies, "He replied cheerfully, 'I did. I did them all' [29, p. 72]."

Discussion

The Polygraph and Voice Stress Analysis

In the formative years of field lie detection, a number of the proponents of the method claimed great success using not a polygraph but merely a measure of one physiological response system. As examples, Marston [30] advocated the use of a "systolic blood pressure test," Benussi [31] a test based on respiratory patterns, and Summers [31] a test based on a measure of electrodermal response (GSR). Although it has been demonstrated today that each of those response systems is useful in detecting deception [32,33], it has also been shown that each makes a separate and independent contribution to the process of lie detection [32,33]. Thus, the polygraph, which simultaneously monitors a number of physiological systems, represents a technological advance over the devices used earlier. It is also certain, however, that as important as it is to record a number of response systems, the manner in which polygraph testing is administered and the way in which polygraphic data are interpreted are of at least equal importance [34,35]. It is recognized today that lie detection is a difficult, complex, and subtle process in which the polygraph instrument itself merely provides the foundation for the structure of what is called the polygraph technique, the art, if you will, of detecting deception with a polygraph instrument. Hence, since the polygraph itself does not detect lies, the technique is not infallible. There is, not surprisingly, considerable controversy about how valid (accurate) the technique is; nonetheless, even the most severe critics acknowledge that the evidence clearly shows an accuracy sufficient to justify the use of polygraph testing for certain purposes [36].

Voice stress analysis, according to its proponents, promises a technologically advanced, simple, easy, almost infallible method of detecting deception, that uses, moreover, information collected from only one response system, the voice [25]. Thus, some of the claims made about voice stress devices are not entirely dissimilar to those made in the formative years of polygraphic instrumentation. There, however, the similarity ends. There is no compelling evidence that any voice stress device actually detects a signal (physiological change) that is clearly and dependably related to stress resulting from deception or any other cause. In fact, the reliable evidence that does exist shows that there is no relationship between what the voice stress devices detect and deception-induced stress. Given those facts, by the way, federal and state courts and state regulatory agencies [27, 37, 38] have recently ruled against proponents of the voice stress devices who have sought the same recognition afforded those who use polygraphic instrumentation.

The developmental history of the polygraph technique shows a conscious, continuing concern with standards of selection and training of polygraph examiners, in clear recognition of the fact that the technique is a complex endeavor in which the polygraph instrument plays a necessary but relatively subordinate role to the technique itself [29-32]. Voice stress proponents deny that

A more telling point in response to this objection, however, is that regardless of how data are scored there is very little agreement among raters on voice stress responses. Correlation coefficients in the reported studies are generally quite small and strongly suggest, as Brenner et al. have reported, that "unreliability is built into the basic scoring procedure and [is] not simply a reflection of inexperience [12, p. 352]." In fact, there are at least two reports that show that untrained or inexperienced evaluators agree as often as, if not more often than, experienced evaluators, although neither judge response data very accurately. In one of these reports, an unpublished study by Worth and Lewis,[3] it was found that an untrained evaluator had higher detection rates in a laboratory situation than a trained evaluator, 58% versus 50% where chance expectancy was 0.25. In the second, more recent report it was found that a manufacturer's employee who trains voice stress operators did considerably worse (less agreement with a polygraph-based criterion) in analyzing response data from real life situations than did two trained but less experienced evaluators. In no case did the three raters agree in even 50% of their evaluations, the agreement rate between the two inexperienced evaluators being only 32% whereas the agreement rate between the employee and the other two evaluators averaged only 40% [27]. These low rates of agreement, of course, merely reinforce the various findings showing the low validity obtained with the voice stress devices.

Controlled Field Studies

Two reliable, independent studies deserve special mention at this point. Each of these studies involved an evaluation of the PSE in field situations (criminal testing); therefore, there can be no objection to them on the ground that they were carried out in an artificial setting. Furthermore, in one of these studies the response data were evaluated by three different persons, all of whom were certified as competent analysts by the manufacturer. Thus, there is little doubt that the data were evaluated in a manner consistent with the manufacturer's guidelines.

In the first study, reported by Barland [20], 66 criminal suspects were tested using both polygraphic and PSE equipment. There was no significant relationship between the scores derived from analysis of polygraphic data and those derived from analysis of PSE data. More important, Barland assessed the accuracy of his PSE-based decisions by three criteria: confessions or guilty pleas in court, decisions made by a panel of legal experts on the basis of written documentation in each case, and the outcome in each case in which there was an independent judicial decision made. Barland's results showed that the accuracy of the PSE was not significantly greater than chance expectancy (0.50). Regardless of which of the three criteria was used as the standard of ground truth, the accuracy in each instance averaged about 50%.

A more recent field-based evaluation of the PSE was carried out by the Department of Commerce in Virginia. In that study, the Department of Commerce, the Virginia State Police, and Dektor, Inc., agreed on the design of a study in which blind PSE evaluations were compared to results obtained in polygraph examinations of persons involved in actual criminal investigations. For that purpose Dektor, Inc., trained and certified two operators. Those two operators and an employee of Dektor independently analyzed PSE data in 40 cases in which complete data were available. When the PSE results in those 40 cases were compared to the polygraph-based outcomes there was no significant association between the conclusions reached by the two methods; the PSE results agreed with the polygraph outcomes, on the average, in 39% of the cases, compared to the 33% agreement that would be expected by chance [27, p. 16]. It is of some interest to note also that the PSE operators performed slightly better when their results included PSE data that they claimed were "unusable" than when those data were excluded and that "substantially the worst performance was recorded by the Dektor employee" [27, p. 17].

[1] J. Worth and B. Lewis, "An Early Validation Study with the Psychological Stress Evaluator (PSE)." unpublished paper, Washington and Lee University, Lexington, VA, 1972.

Thus, the authors of this report conclude that "by all conventional standards of proof we have to regard the validity and reliability of the Psychological Stress Evaluator as unproven. Indeed, it appears that by and large its validity and reliability are not only unproven, but rather are disproven [27, p. 19]."

Analysis of the Reports of Voice Stress Users

Although all of the reliable, independent studies have shown consistent results—whether they were laboratory- or field-based—there are several other reports that, according to the proponents of the voice stress devices, support their claims for the effectiveness of those devices in detecting deception. None of these reports, however, meets generally accepted standards within the scientific community, for that reason alone they are of dubious value. Nonetheless, because these reports are the only ones that buttress the proponents' case they will be briefly discussed here.

In 1972, M. Kradz, in an unpublished paper,[4] reported that he had carried out both polygraph and PSE testing simultaneously on 42 criminal suspects; one additional suspect was tested with the PSE only. Of the 43 suspects tested, 27 were said to be "cleared of suspicion" on the basis of the PSE testing; 21 of those were corroborated as innocent by "independent investigation." Of the 16 suspects "not cleared" by the PSE the guilt of each was said to be established by additional investigation or confession or both. Kradz claimed that his results showed that "100% accuracy was produced in those 36 subject examinations for which complete and concrete corroboration was, or later became, available."

Unfortunately, Kradz's report did not reveal a number of details about his method that are critical to a determination of what his findings might actually suggest. For instance, it was not indicated precisely how the actual guilt or innocence (ground truth) was established for each of the suspects, nor was it clear who carried out the "independent investigation" that apparently established the ground truth criterion Kradz used. When asked how he hud ascertained ground truth in his study. Kradz. testified that he used "independent physical evidence" such as "fingerprints, finding of the weapon, the deceased, stolen property, and questioned documents [26. p. 196|." And, when asked if he had used confessions to establish ground truth, Kradz replied: "Oh, no, not even eyewitnesses [26. p. 197]," although his written report states that an "admission of guilt" was used to corroborate guilt in 13 of 16 cases in which suspects were "not cleared." In fact, according to the written report, in 25% of those cases an "admission of guilt" was made before the "independent" investigation. Kradz further testified that he did not use the outcome of trials in which the suspects were involved because "In two cases we disagreed with that [26, p. 197]." He said he himself determined when the evidence was sufficient to establish that the "PSE was worthy of use in criminal justice [26, p. 200}." The latter statement suggests that the independence of Kradz's "independent investigations" is questionable.

Although Kradz has not yet clarified the details of his method,[5] another version of his report,[6] which is distributed as the original study "reproduced verbatim in its entirety," further confounds the issues. This report describes a method and a number of critical details that are different from what the original report described. The second version, for instance, reports that an unspecified number of the PSE charts were evaluated "in the blind," whereas the first version pointed out that both the subject and the examiner (Kradz) discussed during the testing what was indicated on the PSE charts. Moreover, the second version is even less clear about how ground truth was established than was the first version.

[4] M. KracU. "Psychological Stress Evaluator: A Study," first version of an unpublished paper distributed by Dektor. Inc., Springfield. VA. dated 1972.

[5] Personal communication with M. Kradz, Dektor, Inc., Springfield, VA, 24 Jan. I980, 20 Feb. 1980, and 10 March 1980.

[6] M. Kradz, "Psychological Stress Evaluator: A Study," second version of an unpublished paper distributed by Dektor, Inc.. Springfield, VA, dated 1971.

the accuracy of the PSE in detecting cards concealed by 19 criminal suspects who were undergoing polygraph examinations. In those presumably more motivating circumstances, Nachshon and Feldman found that the PSE yielded an average accuracy of 19%, ranging between 15% and 26% for the three evaluators; the PSE did not produce an accuracy greater than chance expectancy (0.20).

Two other laboratory-based studies of the accuracy of voice stress analysis were reported by Horvath [*22,23*] at Michigan State University. In the first study, 60 college students, 30 male and 30 female, attempted to conceal numbered cards chosen from a deck of five cards while undergoing simultaneous PSE and polygraph testing. Analysis of PSE response data and polygraphic response data, the galvanic skin response (GSR) in particular, was carried out by two trained evaluators. The detection rates obtained with the PSE averaged 22.5% against chance expectancy of 0.20 and were not significantly affected by subjects' sex, repeated trials of testing, simultaneous use of polygraphic and voice stress equipment, or differences between the two trained evaluators of the PSE data. In that same study, detection rates obtained in scoring GSR responses averaged 68.6% (in the first trial of testing only) against chance expectancy of 0.20, and in all cases the rates were significantly greater than chance.

Horvath [*23*] also investigated whether or not the accuracy of the PSE could be enhanced by increasing the subjects' motivation to deceive. In this study 64 college students were promised a reward for successfully completing a task involving the concealment of a numbered card chosen from a deck. In spite of the evidence showing that the subjects were indeed considerably motivated by the reward, that motivation did not increase detection rates obtained with voice stress analysis beyond chance levels; the PSE averaged only 18% correct detections against chance expectancy of 0.20. On the other hand, detection rates obtained with only the GSR in that same study averaged 52%, significantly exceeding chance levels.

It is of some interest to note that in both of the studies reported by Horvath, voice stress analysis yielded lower detection rates than were obtained by analysis of each of the three physiological measures recorded polygraphically—GSR, respiration, and cardiovascular activity [*24*]. Thus, Horvath's findings were remarkably consistent with those reported by Kubis [*18*]; when evaluated in similar contexts voice stress analysis did not yield an accuracy similar to that obtained with the polygraph.

In a recently reported study, Brenner et al [*12*] carried out a lie detection task in which the PSE was used to 'detect ten items of personal information concealed by 20 college students. The students were offered a reward if they were successful in avoiding detection of the items. By random scoring of the subject's PSE responses, an average of 20% of the concealed items would have been detected. The results of the analysis showed that the actual detection rates were not significantly different from chance levels. Depending on the manner in which the PSE responses were scored the detection rates varied between 18.6 and 21.0%. When only the clearest voice stress charts were separately evaluated detection rates remained at chance levels; in spite of the large variation noted in the nature of the stress responses, the variation was not related to the experimental manipulations. Brenner et al point out, moreover, that when used to detect concealed information in the same manner as they used the PSE, the polygraph has yielded detection rates as high as 100%.

Objections to the Controlled Studies

The studies discussed to this point represent the bulk of the reliable evidence reported to date about the effectiveness of voice stress analyzers in detecting deception. Although that evidence clearly does not support the claims made about voice stress analyzers, the proponents of such devices challenge that evidence on two major grounds. First and perhaps foremost among the proponents' arguments is that most of the reliable evidence has been laboratory-based and has involved mere "game playing" situations with low levels of jeopardy. Since, they say, the devices were not designed to be used in such situations it is not surprising that they would be found to be ineffective in them. Although this argument has some ostensible merit, there are a number of points made by the proponents themselves that mitigate its authority.

According to the manufacturers and proponents of the voice stress devices they have the capability to detect absolute stress levels [3-5, 25]. Presumably, such a claim suggests that not only can the devices detect whether stress is present but also the degree of stress, a claim, by the way, which is a significant feature of the training program of the manufacturers. If, of course, this claim were true, then whether or not the testing situation involved low or high levels of stress would generally be of little consequence; that is, if the devices did detect absolute stress levels one would expect to be able to determine easily, for instance, which of a group of items yielded the greatest degree of stress. The evidence does not suggest that possibility.

On the other hand, if there is a certain degree of jeopardy (stress) necessary to obtain valid results with the voice stress devices, as the proponents also claim, what is the threshold and what is the criterion by which one determines it? Is it always present in real life and never in laboratory situations? Those issues have not yet been addressed by the proponents, nor is there any information given about them in training manuals and other material offered by the proponents [25].[2]

In explaining how the prototypical voice stress device was developed, its coinventor has stated that "We set up a known stress/nonstress situation on tape and ran experimental charts with various types of signal processing to attempt to detect any change which may occur which was notable in the stress area which would differ from the representation in the unstressed area and, proceeding with this into refinement, we were able to increase the effectiveness of this by changes in signal processing [26, p. 111]." That testimony about how the first voice stress device was developed and perfected appears to be at odds with the proponents' claim that the device was not designed to detect stress in experimental situations. In that same testimony, in response to a question about what validation studies were done, it was further stated that there was "extensive use of the 'To Tell the Truth' program as broadcast over television, simply because it provided us with a difficult situation where jeopardy of the usual type of lie detection jeopardy was not present. We had singular success with this [26, p. 112].... I think the To Tell the Truth' [accuracy] was something like 94.7 percent [26, p. 145]." At a later point in the testimony addressing the validity of the device in situations where there is less than real-life jeopardy, it was stated that "the PSE doesn't do particularly well in this unless the individual is specifically trained for that application. Our salesmen can do it. The usual PSE examiner is not taught to do that. That is not what they're using it for [26, p. 129]."

The inconsistency between the claims for and about voice stress devices, and the proponents' major abjection to the laboratory-based studies, is obvious. On the one hand, the devices were not designed to be used in experimental situations; on the other, that is precisely how they were developed and validated. On the one hand, the devices are not effective in experimental situations because the stress levels are too low: on the other hand, it is not the devices that are at fault here, since salesmen can apparently be taught how to detect low-jeopardy lies. Thus, it is far from clear why those who have been trained to actually use voice stress devices in detecting deception have been unable to demonstrate their validity in controlled situations.

A second objection made to the studies about voice stress devices is that the operators in those studies did not use valid chart reading techniques, that is, that they did not analyze the response data in a proper manner [25]. This objection, like the one already discussed, does not square with the evidence. In each of the lie detection studies discussed previously [12,18-24] the response data were analyzed by evaluators trained and certified by a major manufacturer as being qualified to interpret data. Moreover, it is clearly expressed in those studies that the criteria advocated by the manufacturer were indeed those that were applied in analyzing the data.

[2] Personal notes from PSE training course, sponsored by Dektor, Inc., Springfield VA, December 1975. See also Refs 2 and 4..

Since the development of the PSE a number of other voice stress analyzers have appeared on the market. According to the advertising literature about these devices they also detect a subaudible microtremor in the voice; thus, the theoretical physiological basis for these devices is identical to that claimed for the PSE. Some of them, however, are engineered so as to obviate the need for a graph-recorded display. Instead, they produce a direct, instantaneous analysis of the voice microtremor and signal "stress" by means of a series of flashing lights or a digital readout [3-5].

Although voice stress analyzers have other obvious applications, they are primarily marketed as a technological breakthrough in the field of lie detection. Because contactual sensors are not necessary and because a subject need be neither present nor even aware that he or she is undergoing a lie detector test, the voice stress devices are reported to be more versatile than, yet as effective as, the traditional polygraph. In fact, it is the purported versatility of those devices and their apparent usefulness in noncontemporary and covert situations that have captured the imagination of the popular media; for example, it has been reported by proponents of voice stress analysis that Lee Oswald was truthful in his denial of shooting President Kennedy [6], that President Carter lied about Bert Lance, and that Ted Kennedy told the truth about Chappaquidick [7].

The purpose of this paper is to discuss and analyze the major empirical evidence pertaining to the claims made about voice stress analysis, in particular, the assertion that voice stress devices are effective in lie detection. Because voice stress devices are usually compared, to the polygraph in the research literature (as well as in advertising literature for the voice devices) a limited companion of results obtained with those two instruments will be made. Before discussing that research, however, it will be useful to discuss briefly some of the other claims made about voice stress analyzers.

The Microtremor Theory

Voice stress analyzers are said to detect subaudible, low-frequency modulations to the 8 to 12 Hz range in the voice. There have been several acceptable attempts to test that claim. Shipp and McGlone [*8*] found no etectromyographic evidence of low-frequency tremors in the laryngeal muscles in the vocalization of either truthful or deceptive utterances. Similarly, McGlone and Hollien [*9*] spectrographically analyzed speech samples of subjects who read a passage in an unstressed condition and of subjects who read a passage while receiving a jerks of electrical shocks; they found no low-frequency energy in the speech samples of either group of subjects. Inbar and Eden [*10*], however, have reported that their research, in which electromyographic recordings were correlated with frequency changes in the voice spectrum, does suggest the existence of low-frequency voice tremors generated by the central nervous system. Thus, the evidence supporting the premise on which the voice stress analyzers are based is not well developed and is certainly not compelling. Nonetheless, even if the microtremor explanation is incorrect, that would not necessarily imply that the devices do not detect some vocal manifestation related to emotional stress.

Detection of Stress

There have been a number of studies carried out to determine the relationship between what the voice stress analyzers detect and accepted traditional indicators of emotional stress. Many of these studies were well-controlled, reliable assessments; the results, however, have been mixed. VanDercar et al [*11*], for instance, reported that they were unable to replicate their own findings of a relationship between PSE voice stress measures and heart rate and A-State scores from the State Trait Anxiety Inventory. Similarly, Brenner et al [*12*] were unable to obtain consistent results with a PSE voice stress analyzer in detecting stress caused by deception and that caused by performance of mental arithmetic tasks; the latter was related to voice stress patterns whereas the former was not. Lynch and Henry [*13*] found that PSE voice stress patterns were not effective in the identification of either stressful or un-stressful words spoken by 43 college students. On the other hand, Borgen and Goodman [*14*] found systematic changes in PSE voice stress measures with the Stroop color/word conflict task; those changes appeared to accompany changes in other psychophysiological measures. Other investigators have also reported a relationship between voice stress measures and indicators of stress, particularly when self-reports of subjects are the criteria [*12, 14-17*]. Thus, the available literatuare does not demonstrate that voice stress analyzers clearly and unfailingly detect emotional stress; the research results have been very inconsistent and the issue needs much more research before it will be settled. It is possible, furthermore, that such research may show the voice microtremor to be a voluntarily controlled component of the voice that is related to stress and anxiety in a largely unpredictable way; the reports of Inbar and Eden [*10*], VanDercar et al [*11*], and Brenner et al [*12*] suggest such an outcome.

Detection of Deception: Controlled Studies

Unlike the research reported pertaining to other claims made about voice stress analyzers, the well-controlled studies in which lie detection has been at issue have yielded consistent results: none of them has shown that the devices are effective in detecting deception. Because there are relatively few of these studies, they will all be discussed here briefly.

The first scientifically acceptable study of the validity of voice stress devices in lie detection was reported by Kubis [*18*] at Fordham University in 1973, about three years after the prototypical instrument was marketed as a lie detector. Kubis designed an elaborate series of studies to determine the relative effectiveness of the polygraph, the PSE, and another voice stress device, the Voice Stress Analyzer (VSA), produced by Decision Control, Inc., in detection of deception. Kubis's study consisted of a "mock crime paradigm" in which some college students were assigned the role of thief, some were the lookout, and some the innocent bystander. Kubis's findings showed that neither the PSE nor the VSA was effective in discriminating between the three student roles. The PSE yielded an accuracy of 32% (27/85) in detecting individual's roles in one portion of that study and 38% (24/63) in detecting roles within each three-student grouping in another portion, against chance expectancy of 33% in each case; the VSA showed an average accuracy of 36% (39/108) in those same situations. On the other hand, polygraphic analysis in Kubis's experiment showed a highly significant overall detection rate of 76%. It is of some interest to note here that Kubis also found that the conditions of his study were sufficiently motivating to produce observable behavioral differences between truthful and deceptive subjects; persons who evaluated only the subjects' behavior during testing were able to discriminate between truthful and untruthful subjects with greater accuracy (53%) than was obtained with the PSE or the VSA.

In another study, Barland [*19*] carried out two small-scale projects to determine the accuracy of the PSE in lie detection. In the first, he had a group of 16 college students conceal information; they were then tested with the PSE to determine if the concealed information could be detected. The results of that experiment showed that the accuracy of the PSE was at chance levels, 6.25% (1/16), a finding that Barland believed to be related to the students' lack of motivation to deceive. To investigate that hypothesis, Barland, in his second project, tested 14 actual criminal suspects—believed to be highly motivated to deceive—with the PSE and the polygraph. He reported initially that the PSE appeared to indicate reliable changes in the voice associated with deception and that the PSE was more effective in conditions of heightened motivation. In another study, larger in scale and more carefully executed, however Barland [*20*] found that the accuracy of the PSE (averaging 51%) did not exceed chance levels (0.50) in detecting deception in criminal suspects, whereas in the same circumstances the polygraph yielded an accuracy of about 90%. Thus, Barland's original hypothesis about the effect of motivation on the effectiveness of voice stress analysis was not supported in his own research.

Nachshon and Feldman [*21*] reported a series of studies designed to investigate the effectiveness of voice stress analysis in detecting concealed information. In one portion of their study, 20 college students concealed cards chosen from a deck of six cards. The students were then tested with the PSE; evaluation of the PSE data by three trained evaluaton yielded an average accuracy rate of 30%, a result not significantly greater than chance expectancy. In another portion of their study, Nachshon and Feldman evaluated

Shipp, T, & Izdebski, K. (1981). Current evidence for the existence of laryngeal macrotremor and microtremor. *Journal of Forensic Sciences*, 26. 501-505.

Smith, G.A. (1977). Voice analysis for the measurement of anxiety. *British Journal of Medical Psychology*, 50, 367-373.

Tippett, R.G. (1995). Comparative analysis study of the CVS A and the polygraph. NITV *Journal of Continuing Education*, First Half 1995, 9-26.

VanDercar, D.H., Greaner, J., Hibler, N.S., Spielberger, CD., & Block, S. (1980). A description and analysis of the operation and validity of the psychological stress evaluator. *Journal of Forensic Sciences*, 25, 174-188.

Detecting Deception: The Promise and the Reality of Voice Stress Analysis

Frank Horvath

Abstract

Within the past decade, a number of so-called voice stress analyzers have been marketed for law enforcement and forensic science purposes. These devices are said to extract from the vocal spectrum a subaudible microtremor signal that is useful in detecting stress in a speaker's voice; thus, it is claimed these devices have great utility as lie detectors and are as accurate as the traditional polygraph instrument. A review of the evidence now accumulated about these devices shows that the evidence for the existence of a microtremor in the voice is problematic and that the capability of these devices in detecting stress is equally questionable. Without exception, however, the scientific evidence reported to date shows that voice stress analyzers are not effective in detecting deception: none of these devices has yet been shown to yield detection rates above chance levels in controlled situations- A brief companion of voice stress analysis and polygraphic testing as methods of lie detection is made.

Keywords

criminalistics, lie detection, voice analysis

In the lie detection field, the most widely publicized development in the past decade has been the so-called voice stress analyzer. In advertisements in popular magazines and in various trade and professional journals voice stress analyzers have been marketed as "truth machines"—devices capable of detecting lies with an accuracy that equals or exceeds that of the more traditional polygraph.

There are now some four or five different voice stress analyzers on the market. The prorotypical instrument, and the one most prominently advertised, is the Psychological Stress Evaluator (PSE). The PSE was first marketed in 1971 by two former military intelligence officers who reportedly developed the device for the purpose of carrying out "lie detection" tests in a covert manner, or at least in a manner that did not require attached sensors. According to its manufacturer, Dektor, Inc., the PSE detects and measures subaudible and involuntary frequency modulations (FM) that are superimposed on audible voice frequencies. The frequency modulations, whose strength and pattern are inversely related to the degree of stress in a speaker at the moment of utterance, are said to result from minute oscillations of the muscles of the voice mechanism. Such oscillations, known as physiological tremors [1], are believed to be under control of the central nervous system during nonstressful periods. As stress is imposed, however, the autonomic nervous system gains dominance, resulting in a suppression of the microtremor. This suppression, indicative of emotional stress, is displayed by the PSE as a characteristic blocked or rectangular wave form.

The PSE processes voice frequencies preserved on a normal tape recording, using electronic filtering and frequency discrimination techniques. The stress-related FM patterns, displayed on a moving strip of heat-sensitive paper, can be processed in four different modes of display for either gross or detailed analysis; because the recovery of the FM indicator spontaneously occurs with the removal of the stressing stimulus, stress in either narrative or monosyllabic speech can be evaluated [2].

Reprinted with permission from the Journal of Forensic Sciences, Vol. 27, No. 2, Copyright ASTM international, 100 Barr Harbor Drive, West Conshohocken, PA 19428.

previously selected number. This lack of jeopardy may have contributed to the fact that the CVSA instrument and procedures obtained an accuracy which was not significantly different from chance.

Interpretation of the CVSA charts was consistent among evaluators, as evidenced by the high interrater reliability. Surprisingly, the CVSA accuracy results were comparable to those obtained in an earlier study (Cestaro & Dollins, 1996) using a similar numbers test paradigm. In that study, pitch and energy extraction techniques yielded an accuracy of 37% in a numbers test paradigm where chance level was 20%. Although the current study also showed that there may be a predictable relationship between measures of a voice component and stress, however, weak, that relationship is not well understood. There is conflicting evidence related to the laryngeal microtremor hypothesis (Shipp & Izdebski, 1981; Smith, 1977). Even if that relationship were well established, it can only be indirectly assessed by examination of speech patterns, and the patterns can be affected by other endogenous or exogenous mediators such as, voice tract pathology, ambient noise, instrument error (see Schoentgen & de Guchteneere, 1991). Increases in the magnitude of a subject's voice microtremor (an unstressed response) may be related to an underlying laryngeal pathology. Additionally, the use of cassette tape recorders for off-line analysis of voice responses for deception detection may be problematic due to distortions introduced by the recording into the measurement of interest (Doherty & Shipp, 1988). This type of analysis has been popularized by the proponents of the CVSA. One or more of these mitigating factors, in concert with a weak stress-inducing laboratory paradigm, can have a serious effect on successful differential diagnoses of response.

The arguments for or against the use of voice stress analysis may ultimately be counterproductive. Such arguments do not consider its potential utility in the arsenal of tools for deception detection. Perhaps investigators should re-examine speech as an additional component rather than to assess it as a singular response channel. Except for the findings of Horvath (1978) related to GSR and PSE, it has not been established how the voice stress channel would perform when compared with each of the three channels currently employed in the traditional polygraph instrument. Atypical differential responding across the three archetypal channels (GSR, pneumograph, and cardiovascular) is common and is largely a function of individual differences in responders.

In summary, the accuracy of examiner decisions concerning subject veracity obtained using the polygraph instrument and procedures was significantly greater than both chance and that obtained using the CVSA instrument. The accuracy of examiner decisions concerning subject veracity obtained using the CVSA instrument and procedures was not significantly greater than chance. While the study design was sufficiently powerful to detect such differences had they existed, subjects did not experience jeopardy during testing—as they would in the field. The lack of jeopardy may have contributed to the obtained relatively low accuracy rates for both instruments. Finally, interrater agreement for the CVSA and polygraph instruments and procedures were both relatively high and significantly better than chance—suggesting that the observed difference in accuracy rates are attributable to instrument/procedure sensitivity—or the lack thereof—rather than examiner test data evaluation skills.

Acknowledgements

The author wishes to express special thanks to the volunteers who participated in the study; Sarah Tidwell and Jeff St. Cyr who assisted throughout the study; David Devine and Special Agent Bob Tippett of the Florida Department of Law Enforcement for administering the psychophysiological detection of deception (PDD) and computer voice stress analyzer (CVSA) examinations; Larry Broadwell, John Schwartz, Dr. Gordon Barland, and Dr. William Yankee for evaluating the PDD tests; Dr. Andrew Dollins for his assistance in manuscript preparation; and Lt. Kelly Vaughan of the Martin County (Florida) Sheriffs Office, Detective Robert Neubauer of the Osceola County (Florida) Sheriffs Office, Don Nettles of Nettles and Associates Investigations, Inc., and Sheriffs Deputy Ronald Cucchiara of the Martin County Sheriffs Office for evaluating the CVSA tests. This project was funded by the Department of Defense Polygraph Institute as DoDPI94-P-0027. The views expressed in this article are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. The results of this study were previously reported by Cestaro (1995). Write for reprints to Dr. Cestaro at Building 3195, Department of Defense Polygraph Institute, Ft. McClellan, AL 36205-5114

References

Barland, G.H. (1974). Use of voice changes in the detection of deception. *Polygraph*, 7, 129-140..

Brenner, M. & Branscomb, H.H. (1979). The psychological stress evaluator; Technical limitations affecting lie detection. *Polygraph*, 8, 127-132.

Brenner, M., Branscomb, H.H., & Schwartz, G. (1979). Psychological stress evaluator—two tests of a vocal measure. *Psychophysiology*, 16. 351-357.

Cestaro, V.L., & Dollins, A.B. (1996). An analysis of voice responses for the detection of deception. *Polygraph*, 25. 15-34.

Cestaro, V.L. (1995). A comparison between decision accuracy rates obtained using the polygraph instrument and the computer voice stress analyzer (CVSA) in the absence of jeopardy (Report No. DoDPI95-R-0002). Fort McClellan, AL: Department of Defense Polygraph Institute.

Doherty, E.T., & Shipp, T. (1988). Tape recorder effects on jitter and shimmer extraction. *Journal of Speech and Hearing Research*, 31. 485-490.

Fleiss, J.L. (1981). Statistical methods for rates and proportions, 2nd ed. New York: John Wiley & Sons.

Gustafson, L.A., & Orne, M.T. (1963). The effects of heightened motivation on the detection of deception. *Journal of Applied Psychology*, 47. 408-411.

Horvath, F.H. (1978). An experimental comparison of the psychological stress evaluator and the galvanic skin response in detection of deception. *Journal of Applied Psychology*, 63. 338-344.

Horvath, F.H. (1979). Effect of different motivational instructions on detection of deception with the psychological stress evaluator and the galvanic skin response. *Journal of Applied Psychology*, 64, 323-330.

Horvath, F.H. (1982). Detecting deception: The promise and the reality of voice stress analysis. *Journal of Forensic Sciences*, 27, 340-351.

Humble, C. (1995). From the director's desk NITV *Journal of Continuing Education*, 12, 1-2.

Lieblich, I., Naftali, G., Shmueli, J., & Kugelmass, S. (1974). Efficiency of GSR detection of information with repeated presentation of series of stimuli in two motivational states. *Journal of Applied Psychology*, 59, 113-115.

Lynch, B.F., & Henry, D.R. (1979). A validity study of the psychological stress evaluator. *Canadian Journal of Behavioral Science*, 11, 89-94.

Schoentgen, J., & de Guchteneere, R. (1991). An algorithm for the measurement of jitter. *Speech Communication*, JO. 533-538.

(1 to 10) to serial position of those numbers in the test series, adjusted for the two padding sequences. Thus, regardless of the starting number in a subject's sequence, all scored numbers would fall into the range 1 through 4, with the key appearing only in position 2 or 3. Each scorer had 1 chance in 4 of correctly identifying the key by chance alone.

Data analyses

Analyses included a test of the significance of the proportionality between correct number determinations and chance accuracy (25%). Effects of examination order on mean accuracy were also examined. A power analysis performed prior to the study indicated that with N-42, power > 0.09 for the expected effect size (0.25). The Fleiss (1981) multiple rater Kappa test was used to independently assess decision agreement among the four evaluators within each instrument category.

Results

Evaluator Accuracy

PDD evaluators correctly identified the correct key number in 105 of 168 (42 subjects x 4 evaluators) total trials, achieving a statistically significant overall accuracy of 62.5% (p < .05), with a range of 57% to 69%. Three of the four evaluators obtained accuracy rates equal to or greater than 60%. The CVSA evaluators correctly identified the correct key number in 65 of 168 total trials, obtaining a nonsignificant overall accuracy of 38.7%, with a range of 24% to 45%. Three of the four evaluators achieved accuracy rates equal to or greater than 40%. The difference (23.8%) between mean accuracy rates obtained using the two instruments and their procedures was statistically significant (p. < .05).

Order Effects on Accuracy

The order of examination administration had an effect on the accuracy of each instrument; accuracy declined on the second series of tests. The PDD mean accuracy obtained using the polygraph instrument was 75% (p_ < .05) for the 21 subjects undergoing the PDD examination before the CVSA examination. The mean accuracy obtained using the polygraph instrument was 50% (p_ < .05) for the 21 subjects undergoing the CVSA examination before the PDD examination. Similarly, when the PDD examination preceded the CVSA tests, overall CVSA accuracy dropped from 41% (p_ > .05) to 35% (p_ > .05). The changes in accuracy rates within each instrument category were not statistically significant. Decision accuracy was not affected by subject gender.

Interrater Reliability

Three out of four PDD blind evaluators agreed on the number selected for 30 of the 42 subjects, with 16 unanimous agreements. The correct number was identified for 24 of those 30 subjects. Three out of four CVSA blind evaluators agreed on the number selected for 31 of the 42 subjects, with 2 unanimous agreements. Fourteen of those 31 subjects' numbers were correctly identified. There were six cases in which both the CVSA and PDD evaluators agreed, with five of the six correctly identifying the subjects' selected numbers.

The frequency of agreements on serial position of the key number among evaluators for each subject was examined using the Kappa statistic for multiple ratings (Fleiss, 1981), the results of which are shown in Table 1. With the exception of position 4, agreement among evaluators was statistically significant for each possible position of the key item, as was overall agreement. However, the key numbers could be physically located only in positions 2 and 3, dependent on question padding.

Table 1

Interrater Agreement (Kappa) Among PDD and CVSA Evaluators

POSITION

Exam	1	2	3	4	Not Scored	Overall
PDD	.26*	.58*	.52*	.02	.10	.46*
CVSA	.65*	.35*	.48*	.20*		.42*

*p < .05

Note: PDD = psychophysiological detection of deception; CVSA = computer voice stress analyzer.

Discussion

Data analysis indicates that, under similar test conditions, the percent of correct subject veracity decisions made using information gathered during a PDD examination exceeded the percent of correct subject veracity decisions made using information gathered during a CVSA examination by 23.8%--a statistically significant difference. These results suggest that, under the test conditions used, although the CVSA instrument performs electrically as theorized (Cestaro, 1995), it has less sensitivity to psychophysiological reactivity than the traditional polygraph instrument. These differences may be a function of the additive information, or Gestalt, provided by the multi-channel structure of the polygraph instrument versus the difficulties imposed by the single channel analysis of the CVSA, particularly when a conflict arises (e.g., none of the responses meet the decision criterion). Although, in certain situations, individual PDD examiners may rely heavily on one of the measures, it is not likely that an experienced examiner will be satisfied with a decision based on that single component. A power analysis conducted prior to beginning the study indicated that the design had a 0.90 probability of correctly detecting an effect of at least 0.25 different from chance if such an effect actually exists. Thus, failure to obtain subject veracity decision accuracy rates significantly greater than chance using the CVSA suggests that, under the test conditions used, there is a probability of at least 0.90 that the CVSA is not sensitive enough to accurately detect effects at a level of at least 0.25 greater than chance accuracy.

Scoring procedures seem to be as consistent for the CVSA as for the polygraph instrument, as suggested by the high interrater reliability for both instruments and associated procedures. However, Horvath (1978) obtained interrater agreements that were greater for one component of the traditional polygraph instrument than for the voice stress analyzer; r = .92 for the GSR and i = .38 for the PSE. This suggests that scoring biases may have played a major role in the interpretation of voice stress responses using the older instrument. Brenner and Branscomb (1979) submit that the problem of scoring subjectivity is serious enough to bring into question any specific legal decisions made regarding PSE results. In this study, the absence of jeopardy, contrived or real, may have contributed to the low accuracy rates obtained using the polygraph and CVSA instruments and procedures. No incentives were offered to subjects to motivate them to act or react in a particular manner. It was not expected that subjects would experience anything other than very low levels of stress when answering untruthfully about a

Apparatus

A polygraph instrument (Lafayette, lafayette, IN, Factfinder Model 76740/76741) was used to record skin resistance, respiratory, and cardiovascular activity on paper charts. A CVSA (NITV, West Palm Beach, FL) was used to record and display voice response data on paper charts. A lapel microphone (Radio Shack, Fort Worth, TX, Model 33-3003) was connected to the audio jack of the CVSA to present voice responses to the instrument. A voice recorder (TEAC, Montebello, CA, Model 134B) was used to collect voice responses for off-line analysis. A lavaliere microphone (Shure, Evanston, IL, Model S70S) was used with the audio recorder to record subjects' verbal responses. A desktop IBM compatible computer was employed to replay questions throughout testing in both portions of the study.

The questions presented to the subjects were digitized and recorded to computer hard disk using a Sound Blaster board (Model 16ASP, Creative Labs, Inc., Milpitas, CA). A parallel port interface, designed and fabricated inhouse, connected to a Radio Shack (Fort Worth, TX) integrated stereo amplifier (Model SA-155) and two Radio Shack speakers (Model Minimus-77) was used to present questions. This system ensured that each question was presented during the PDD examination and the CVSA examination with the same inflection, and at the same volume, each time it was asked. The question presentation software also controlled the TEAC recorder— which was used to record subject response—through an RS-232 interface designed inhouse.

Examiners

A certified PDD examiner, who had administered more than 5000 examinations over a 30-year period, conducted the PDD portion of the study. A second certified PDD examiner, also trained and experienced in CVSA use by NITV, conducted the CVSA portion of the study. This examiner had over one and a half years PDD experience encompassing 150 examinations, and had also administered 450 CVSA examinations during two and a half years. Four additional certified PDD examiners, who were unaware of subjects' number selections, independently evaluated the polygraph tests to determine the number selected by each subject. The CVSA tests were also independently evaluated by four trained and certified CVSA examiners. Two of these examiners were also trained and certified PDD examiners.

Design

The standardized DoDPI acquaintance (stimulation) test, using a Known Solution POT test was employed for subject testing. The 42 subjects were pseudo-randomly assigned to each of two 21 subject groups. One group was tested with the polygraph instrument first, followed by CVSA testing. The other group was tested with the CVSA first, then the polygraph instrument, to counterbalance the order of testing. Within each group, half of the subjects were randomly assigned to Key A (third question is the key question), and half to Key B (the fourth question is the key number). Each subject was tested using six numbers in sequence. No more than three subjects having the same key position were tested consecutively by either examiner. Since the relevant questions for the two examinations were identical, digitized voice was used to present the questions to the subjects. The only difference between the two examination was the inclusion of irrelevant questions in the CVSA test question format. Data from the PDD and CVSA examinations were independently assessed by four additional examiners who were unaware of subjects' key numbers. The dependent measures were the number of correct determinations, and the number of concurrent determinations made using the two instruments and their processes.

Procedures

Upon arrival at the Department of Defense Polygraph Institute (Fort McClellan, AL), each participate was escorted by a member of the research team to a secluded room and asked to read a brief description of the research project. Individuals indicating that they would participate were asked to read and sign an informed consent affidavit. A brief biographical/medical questionnaire was completed, to ensure that the participant was in good health and not currently taking medication which could interfere with the examination results. The subject was then escorted to one of the examination rooms, determined by prior assignment, for testing. The examinations were then administered as described below. When both the PDD and CVSA examinations were completed, the subject was escorted back to the secluded room for subject debriefing, and to read and sign a debriefing statement.

Procedures Common to CVSA and PDD.

The examiner conducted a pretest interview prior to placing the sensors on the subject. During the interview, the subject was told to select a number between 3 and 8, not 3 or 8, and to write the number selected, one to two inches in height, in the middle of a sheet of paper. The examiner was given the key number position for each subject in order to properly pad the sequence of numbers. The examiner then wrote the padding numbers above and below the number written by the subject and placed the sheet of paper on the wall directly in front of the subject. The Key A sequence used two "padding numbers" before and three after the number selected by the subject. The Key B sequence used three padding numbers before and two after the key number. The subjects were instructed to answer "no" to the questions concerning numbers, even if it meant that they would be lying about the number they chose.

PDD procedure

If the CVSA examination was administered prior to the PDD examination, the previously selected number was used again. The POT test was then administered according to the procedures taught at the DoDPI, with the exception that the post-test interview was not conducted. In order to comply with the philosophy of the CVSA examination procedure to avoid situational stress, no psychological set was required, and the post-test would have served no purpose in this study. If the PDD examination was administered before the CVSA examination, the subject proceeded immediately thereafter to the CVSA examination.

CVSA procedure

If the PDD examination was administered prior to the CVSA examination, the previously selected number was used again. The examiner conducted the pretest interview in a manner to remove any situational stress associated with detection of deception examinations. The lavaliere and lapel microphones were then placed on the subject and the CVSA instrument was calibrated for the subject's voice level. The question sequence was in the CVSA POT format, IR-IR-R-IR-R-IR-R-IR-R-IR-R-IR-R-IR (where IR=irrelevant and R=relevant). Irrelevant questions had no connection to the issue at hand, caused no stress of themselves, and were known truth (e.g., "Am I wearing a tie?"). Relevant questions consisted of numbers from the same set used during the PDD exam.

The CVSA examiner conducted two tests. In accordance with NITV procedures, the first test was discarded to avoid scoring data confounded by situational stress. The second test was retained for scoring. All examinations were recorded on audio and video tape for off-line analysis to confirm the live results.

Data Reduction and Analysis

Test evaluation

Each PDD test was independently evaluated by each of four certified PDD examiners. CVSA tests were independently evaluated by each of four certified CVSA examiners, trained by the NITV. PDD test data evaluation consisted of selection of the response showing the most reactivity, according to traditional PDD procedures. CVSA test data evaluation consisted of selection of the response showing the highest percentage of "blocking" (i.e., rectangularity) on the relevant/associated irrelevant, in accordance with accepted CVSA scoring practices. Although data analysis procedures were clear and explicit, evaluators were not given the padding information (two or three padding numbers) in order to avoid the possibility of biased scoring. However, all evaluators were told that there was at least one padding question before and after the key question (i.e., the key number was never in the first or last position in the sequence), leaving only four responses to be scored.

Data reduction

The dependent measures were: the number of times a scorer correctly identified the number selected by a subject and; the frequency that evaluators using different instruments and processes identified a subject as being deceptive to the same question, irrespective of the accuracy of the decision. Data were transformed from number sequence

A Comparison Between Decision Accuracy Rates Obtained Using the Polygraph Instrument and the Computer Voice Stress Analyzer (CVSA) in the Absence of Jeopardy

Victor L. Cestaro

Abstract

This study was designed to evaluate the decision accuracy and agreement rates obtained using the traditional polygraph instrument and the computer voice stress analyzer (CVSA). Forty-two subjects took psychophysiological detection of deception (PDD) examinations administered with the polygraph and CVSA instruments within the context of Peak of Tension (POT) tests. Half of the subjects were tested on the polygraph instrument, then the CVSA instrument. The remaining half were tested using the instruments in the opposite order. PDD and CVSA based POT tests were blind-evaluated by four independent examiners for each instrument. The frequencies of accurate determinations made using each instrument were compared using proportionality tests. The CVSA instrument and associated processes were significantly less accurate than the polygraph instrument and PDD processes tested in similar circumstances (38.7% vs. 62.5%, with chance = 25%). Interrater reliability, assessed using a multiple rater Kappa test, showed that agreement among all blind evaluators within each instrument category was significantly better than chance (r=< .05). These data indicate there may be a systematic and predictable relationship between voice patterns and stress related to deception, and that the differences observed in accuracy rates between the two instruments are attributable to instrument/procedure sensitivity rather than examiner data evaluation skills.

According to Humble (1995), the detection of deception using voice stress analysis is rapidly gaining acceptance and receiving favorable reviews from many law enforcement agencies. Although reviews of the accuracy of voice analyzers, such as the Psychological Stress Evaluator (PSE, Dektor Counterintelligence and Security, Springfield, VA) have been mixed (e.g., Horvath, 1978, 1979, 1982; VanDercar, Greaner, Hibler, Spielberger, & Block, 1980), the acceptance of instruments employing this technique has been facilitated by their ease of use, non-invasiveness, and short training period required for prospective operators.

Surprisingly, no controlled laboratory research has been conducted to test the validity or reliability of the recently developed computer voice stress analyzer (CVSA) instrument or the techniques employed in its use, nor are there any indications that it meets or exceeds the accuracy rates reported for the traditional polygraph instrument. The high accuracy rates claimed by the manufacturer~The National Institute for Truth Verification (NITV)--are based on field data rather than laboratory research, as stated by Dr. Charles Humble, president and founder of the NITV (G. Barland, personal communication, June 12, 1989):

> The CVSA is a computerized voice stress analyzer that is based on the older Psychological Stress Evaluator. As you are aware, the research concerning the validity of the PSE has always been controversial and never accepted in polygraph circles. The validation of the CVSA was accomplished by utilizing (sic) the audio portion of 75 known-conclusion cases. Twenty of these were NDI and 55 were DI. The CVSA correctly called all of the cases for a correlation rate of 100%.
>
> Rather than rely on laboratory studies which I do not feel accurately reflect the validity of either the polygraph or the CVSA, I would refer to field studies which, in my opinion, do.
>
> [Note: NDI = No Deception Indicated, DI = Deception Indicated].

Manufacturers and proponents of voice stress analysis attribute the failure to obtain high accuracy rates in analog (laboratory) studies to the low level of jeopardy in "game playing" laboratory scenarios (Horvath, 1982). Webster's Dictionary defines jeopardy as "exposure to loss, or damage." Individuals submitting to a detection of deception examination outside of the laboratory usually experience some jeopardy in association with the examination while those submitting to a detection of deception examination in the laboratory, a contrived situation, do not. Tippett (1995) indicates that previous testing has shown that artificially induced jeopardy produced only marginal results with the CVSA. He argues that when at least a moderate level of personal jeopardy was perceived by subjects, the CVSA and the polygraph instruments and processes were equally effective in determining truth or deception. In a study conducted by Tippett (1995), 54 subjects who were undergoing mandatory private therapy related to past sex offenses were tested using the CVSA and the polygraph instrument. According to Tippett, "... there was a 100% agreement between the CVSA and the polygraph." He concluded that the CVSA is as effective as the polygraph instrument for detecting deception.

The personal jeopardy requirement can cause uncertainty in testing, as Horvath points out (1982, p. 344) when he asks, " ... if there is a certain degree of jeopardy (stress) necessary to obtain valid results with the voice stress devices, as the proponents also claim, what is the threshold and what is the criterion by which one determines it?" Horvath (1982) also questioned how the PSE was developed and perfected if it could not be tested in experimental situations. Nonetheless, it was found that PSE, the State/Trait Anxiety Index, and heart rate measures covaried and reflected levels of stress in the first portion of a pair of studies conducted by VanDercar, et al. (1980). The failure to validate the PSE in the second study was attributed to lower levels of induced stress. Furthermore, Barland (1974), in a low/high stress study, demonstrated that the PSE achieved high accuracy rates when stress levels were high, but did not do so with low stress levels. However, Lynch and Henry (1979) found no evidence that the PSE could discriminate between stressful and non-stressful responses at greater than chance levels.

In studies where jeopardy has been defined in terms of motivation (e.g., monetary loss or gain) to pass or fail a detection of deception examination, the results have been mixed. Gustafson and Orne (1963) reported that detection rates were greater for subjects who were motivated than for subjects who were not. However, Lieblich, Naftali, Shmueli, and Kugelmass (1974) claimed that motivation had no significant effect on detection rates. Additionally, Horvath (1979) has shown that increased motivation does not improve the detection of deception with the PSE. Brenner, Branscomb, and Schwartz (1979) also question the validity of the PSE in the context of deception detection, although some aspects of the analysis may be valid for the measurement of stress.

This study was designed to evaluate a second generation voice analyzer, the CVSA, in a laboratory test in the absence of jeopardy. The experiment was designed to determine if accuracy rates obtained using the CVSA instrument and procedures differed from those obtained using the traditional polygraph instrument and procedures. The CVSA and the polygraph instrument were used to test subjects within a Peak of Tension (POT) numbers test paradigm. It was expected that a stressed response to the key number could be detected using both the CVSA and polygraph instruments and would be discriminable from all other responses. Evaluators who were not aware of the numbers selected by subjects made decisions regarding subject veracity based on the examination of paper charts collected during test administrations. Accuracy rates obtained by these evaluators were examined. Additionally, interrater agreement was assessed among evaluators within each instrument category.

Method

Subjects

Forty-two subjects recruited from the U.S. Army training command at Fort McClellan and a local civilian contract agency participated in this study. Subjects were 22 males and 20 females between the ages of 19 and 35 years.

Discussion

Results of laboratory tests indicate that the CVSA functions electrically according to frequency modulation detection theory. It was found that discrete changes in the frequency of the input signal caused discrete deflections of the CVSA pen, and that the amplitude of those deflections was proportional to the frequency of the input signal. Increases in the amplitude of the input signal resulted in reduction of the amplitude of extraneous (noise) signals on the CVSA chart tracings. These findings are consistent with the manufacturer's theory of operation. Thus, if there is an inverse relationship between stress and voice microtremor amplitude, and those changes have sufficient signal value to be detected by the CVSA, it should be possible to see pattern changes in the CVSA output under different levels of stress.

Research substantiating the basic underlying theory, by comparing simultaneous vocal tract muscle activity and voice microtremor, has been minimal (e.g., Inbar & Eden, 1976). Additionally, there is limited research supporting the inverse relationship between microtremor and stress (e.g., Smith, 1977), and that relationship was indirectly assessed by examination of speech patterns. Perhaps research in this area should focus on: (1) the existence of laryngeal microtremor as assessed concurrently by EMG and speech pattern analysis; and (2) autonomic mediation of muscle microtremor, particularly laryngeal microtremor. It has been established that autonomic innervation extends primarily to cardiac and smooth muscle tissue (slow response). Some of the striate (fast response) muscle groups in the larynx are innervated by the vagus nerve (cricothyroid, arytenoid), but there is insufficient information available regarding the function of the vagal innervation.

In summary, the CVSA instrument has been shown to detect discrete changes in speech fundamental frequency using laboratory instruments to simulate voice microtremor, confirming NITV's underlying theory of operation. However, these results do not confirm: (1) the existence of voice microtremor; (2) a relationship between microtremor amplitude and psychological or physical levels of stress; (3) a reduction of microtremor amplitude during the act of deception; and (4) that voice microtremor—if it exists—has sufficient signal value to be detected by the CVSA.

Acknowledgements

This project was funded by the Department of Defense Polygraph Institute as DoDPI94-P-0027. These results were previously reported in Cestaro (1995). The views expressed in this article are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. For reprints, write to Dr. Cestaro at DoDPI, Building 3195, Ft. McClellan, Alabama 36205-5114.

References

Brenner, M., Branscomb, H.H., & Schwartz, G. (1979). Psychological stress evaluator—two tests of a vocal measure. Psychophysiology, 16. 351-357.

Cestaro, V.L. (1995). A comparison between decision accuracy rates obtained using the polygraph instrument and the computer voice analyzer (CVSA) in the absence of jeopardy (Report No. DoDPI95-R-0002). Fort McClellan, AL: Department of Defense Polygraph Institute.

Cestaro, V.I., & Dollins, A.B. (1994). An analysis of voice responses for the detection of deception (Report No. DoDPI94-R-001). Fort McClellan, AL: Department of Defense Polygraph Institute.

Gray, H. (1977). The organs of voice and respiration. In T.P. Pick & R. Howden (Eds.), *Gray's anatomy* (pp. 955-983). New York: Gramercy Books.

Inbar, G.F., & Eden, G. (1976). Psychological stress evaluators: EMG correlation with voice tremor. *Biological Cybernetics*, 24. 165-167.

Kahane, J.C. (1986). Anatomy and physiology of the speech mechanism. In H. Halpern (Ed.), The Pro-Ed studies in communicative disorders (pp. 78-93). New York: Pro-Ed.

Lippold, O. (1971). Physiological tremor. Scientific American, 224. 65-73 Motley, M.T. (1974). Acoustic correlates of lies. *Western Speech*, 38, 81-87.

Motley, M.T. (1974). Acoustic Correlates of Lies. *Western Speech*, 38, 81-87.

NITV. (1994). Certified Examiners Course Manual. (Available from the National Institute for Truth Verification, West Palm Beach, FL).

O'Hair, D., & Cody, M.J. (1987). Gender and vocal stress differences during truthful and deceptive information sequences. *Human Relations*, 40. 1-14.

Shipp, T., & Izdebski, K. (1981). Current evidence for the existence of laryngeal macrotremor and microtremor. *Journal of Forensic Sciences*, 26, 501-505.

Smith, G.A. (1977). Voice analysis for the measurement of anxiety. *British Journal of Medical Psychology*, 50, 367-373.

Streeter, L.A., Krauss, R.M., Geller, V., Olson, C., & Apple, W. (1977). Pitch changes during attempted deception. *Journal of Personality and Social Psychology*, 35_(5), 345-350.

Tolkmitt, F.J., & Scherer, K.R. (1986). Effect of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology*, 12. 302-313.

VanDercar, D.H., Greaner, J., Hibler, N.S., Spielberger, C.D., & Block, S. (1980). A description and analysis of the operation and validity of the psychological stress evaluator. *Journal of Forensic Sciences*, 25, 174-188.

Zalewski, J., Majewski, W., & Hollien, H. (1975). Cross-correlation between long-term speech spectra as a criterion for speaker identification. *Acustica*, 34. 20-24.

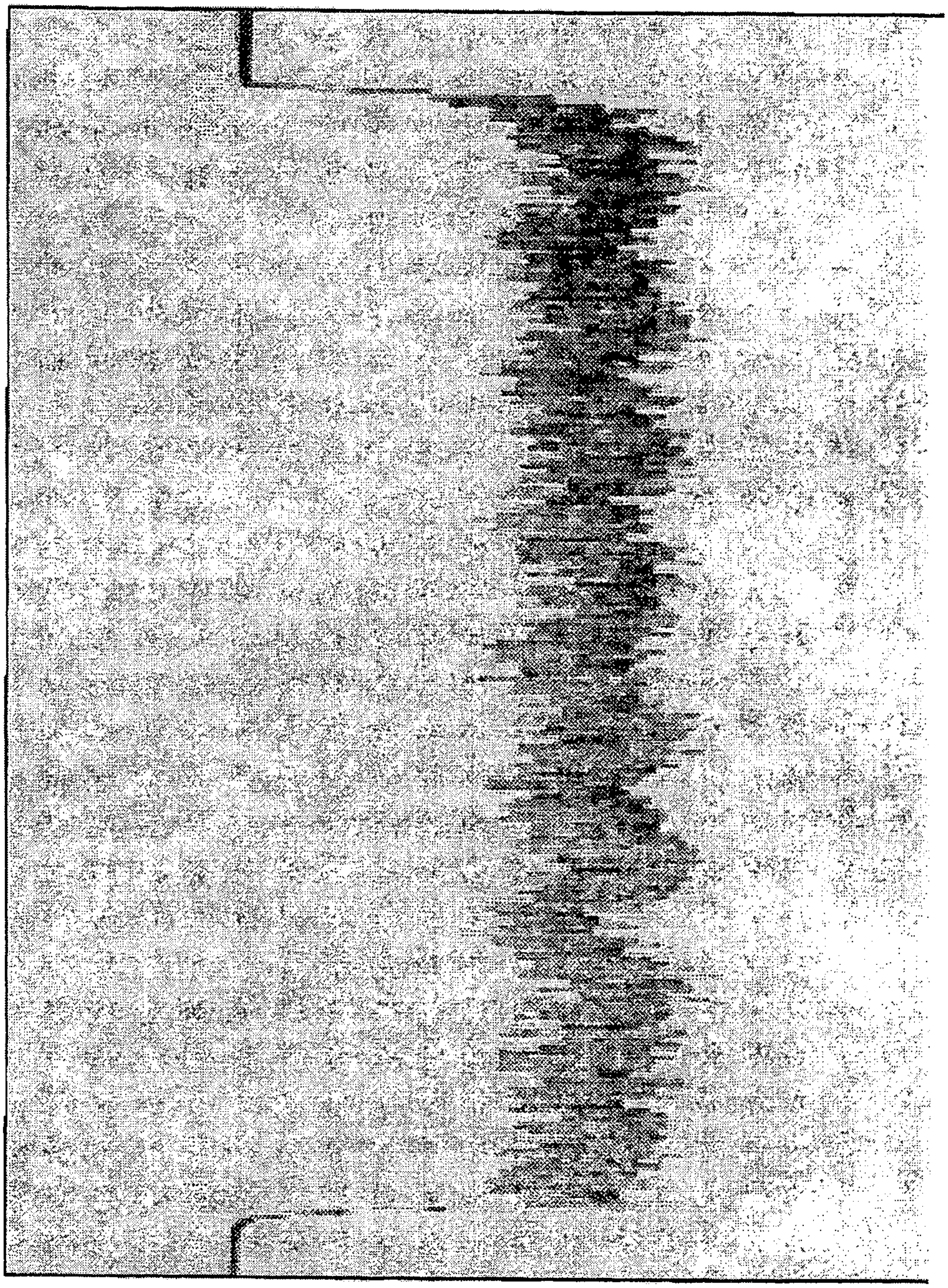


Figure 9. CVSA output when the input is an 8 millivolt peak-to-peak 300 Hz sine wave.

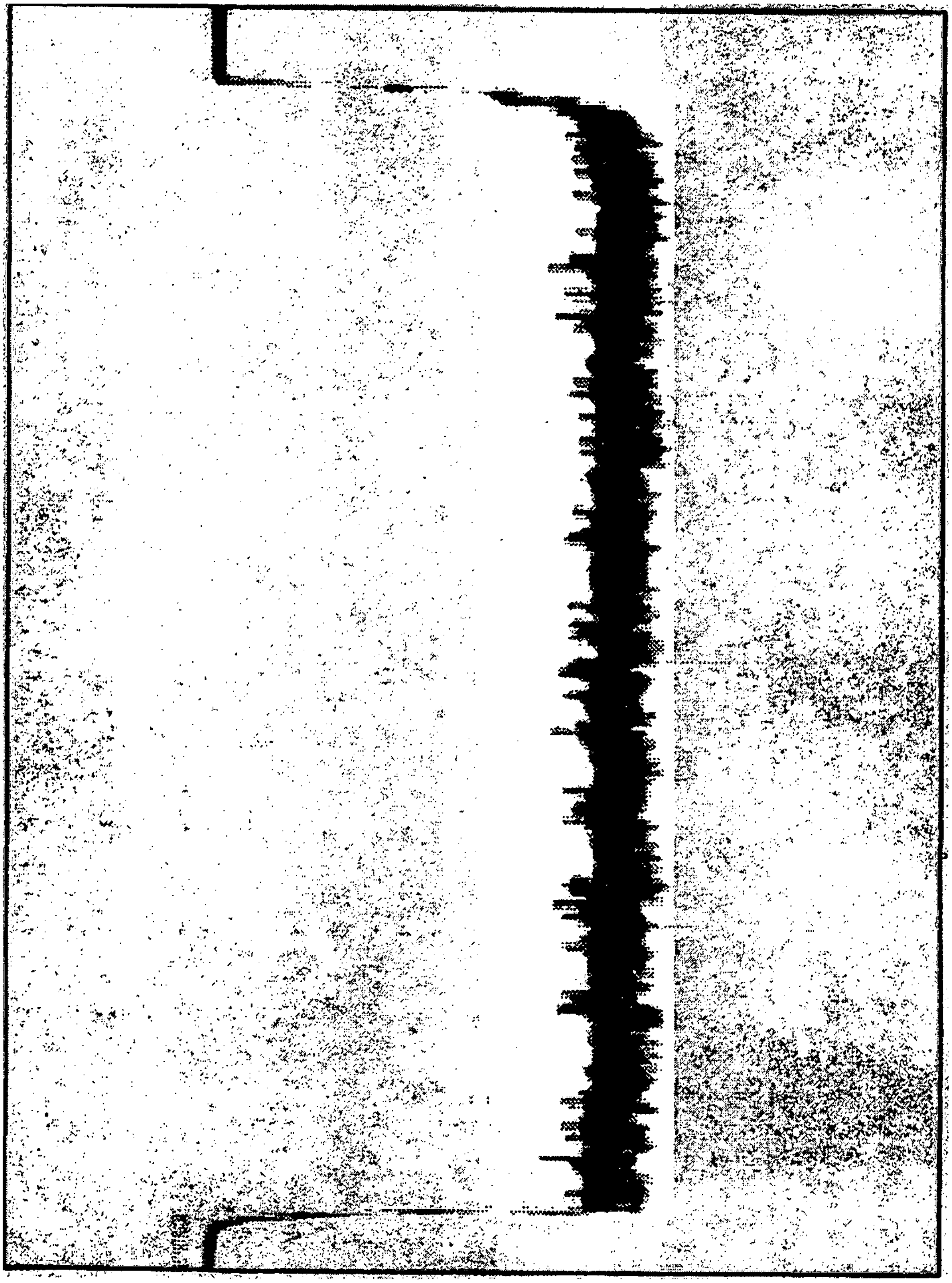


Figure 10. CVSA output when the input is a 32 millivolt peak-to-peak 300 Hz sine wave.

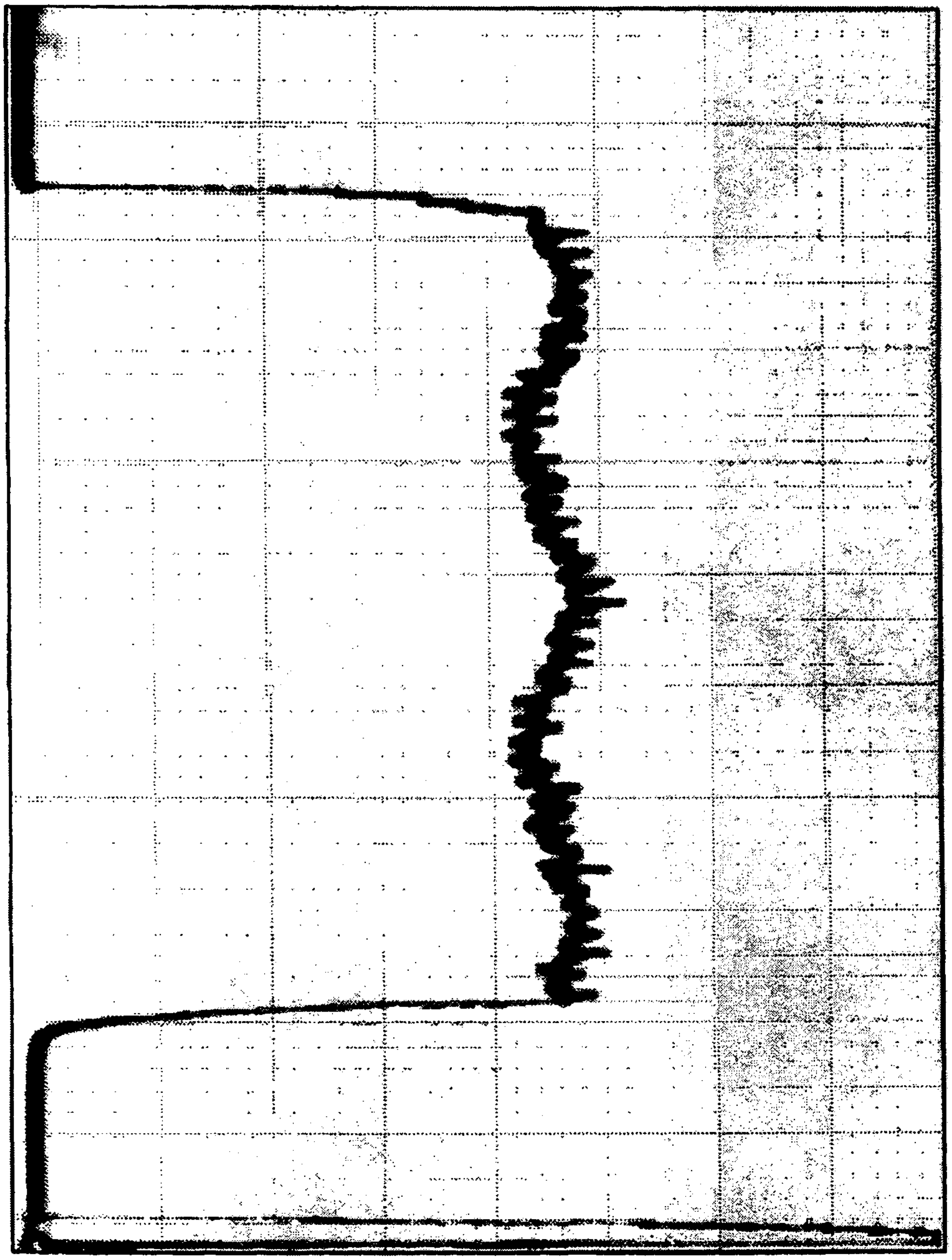


Figure 8. CVSA output when the input is an 8 millivolt peak-to-peak 1000 Hz sine wave modulated at a 10 Hz rate (unstressed response).

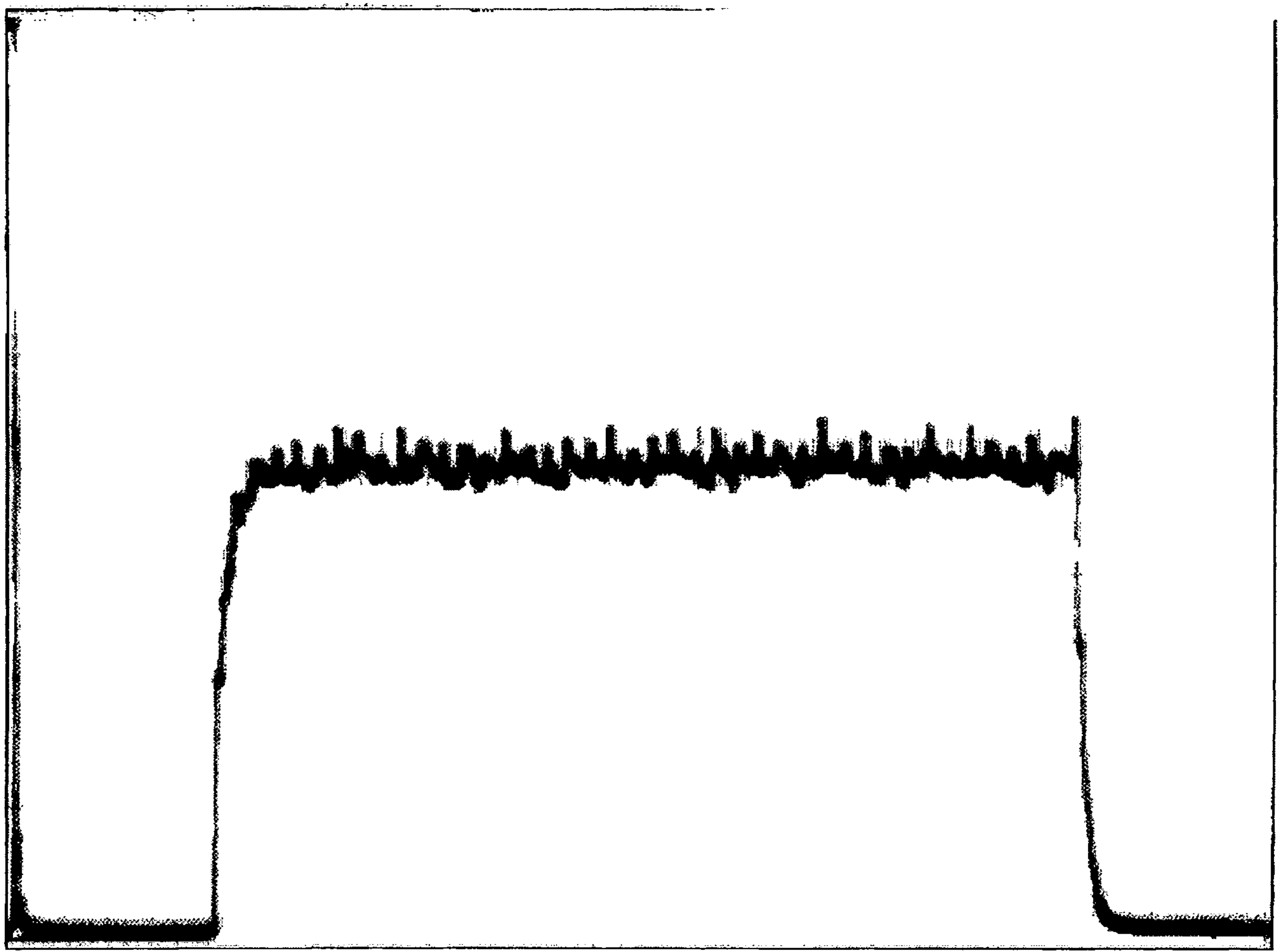


Figure 7. CVSA output when the input is an 8 millivolt peak-to-peak 1000 Hz sine wave (stressed response).

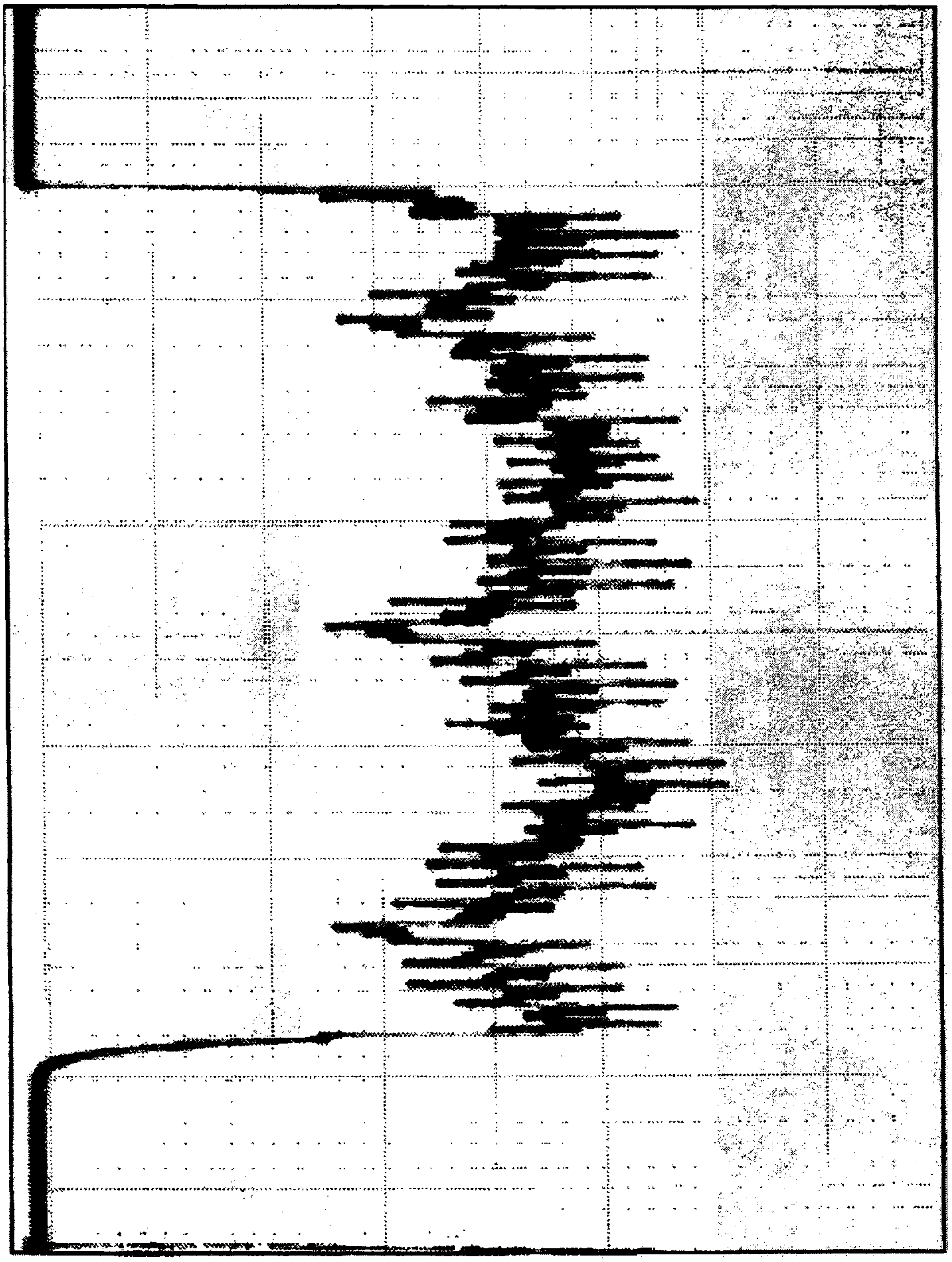


Figure 6. CVSA output when the input is an 8 millivolt peak-to-peak 500 Hz sine wave modulated at a 10 Hz rate (unstressed response).

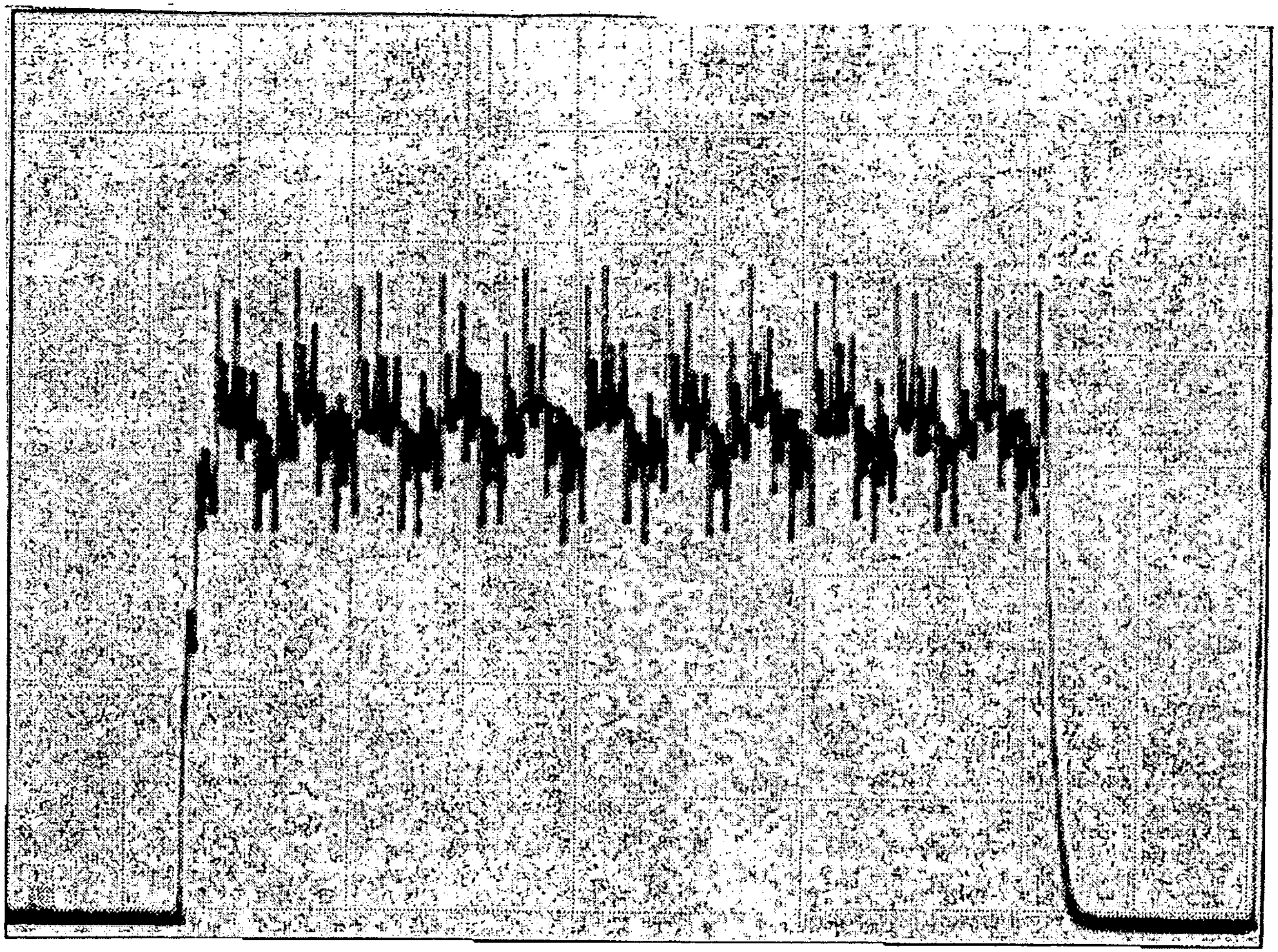


Figure 5. CVSA output when the input is an 8 millivolt peak-to-peak 500 Hz sine wave (stressed response).

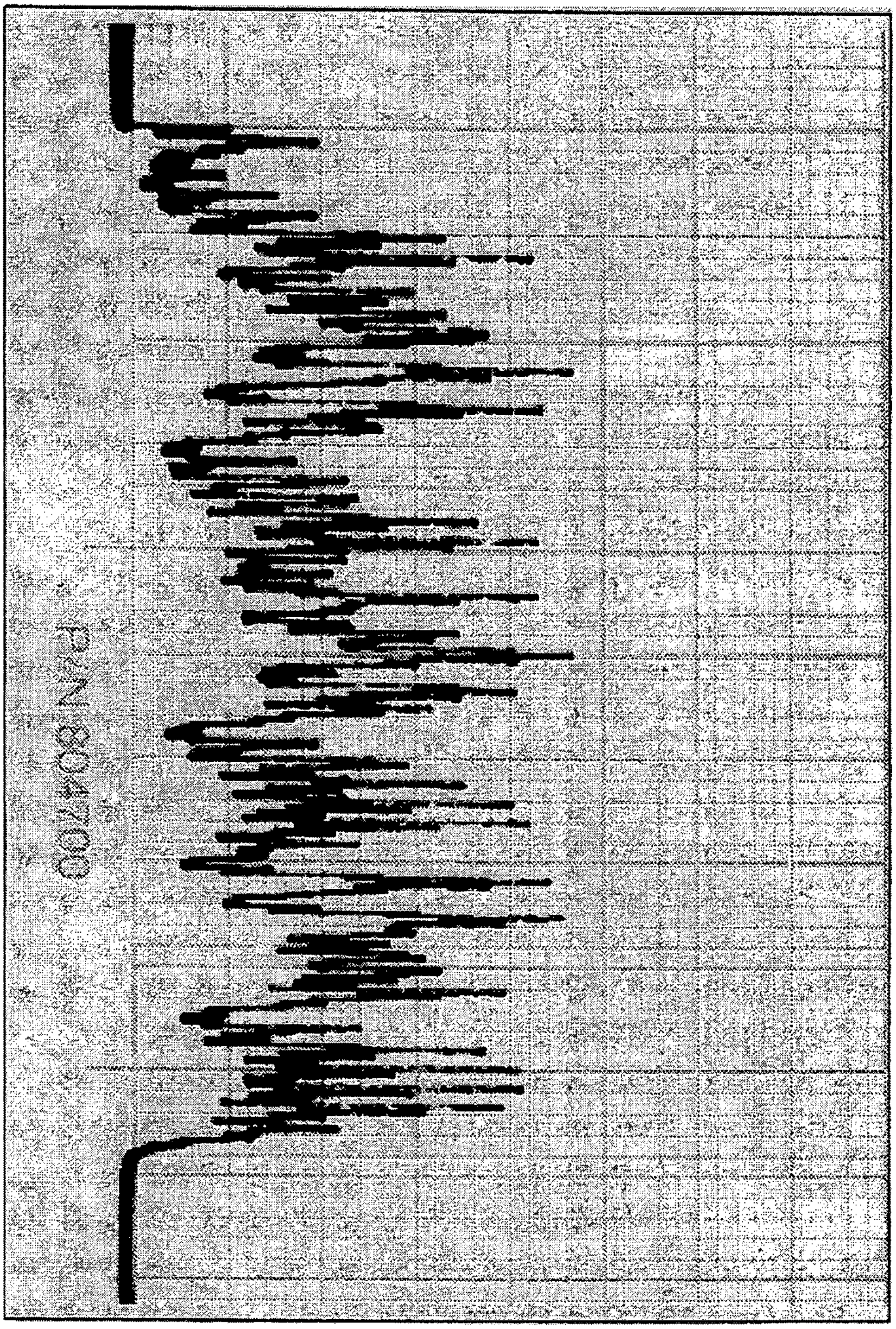


Figure 4. CVSA output when the input is an 8 Millivolt peak-to-peak 200 Hz sine wave modulated at a 10 Hz rate (unstressed response).

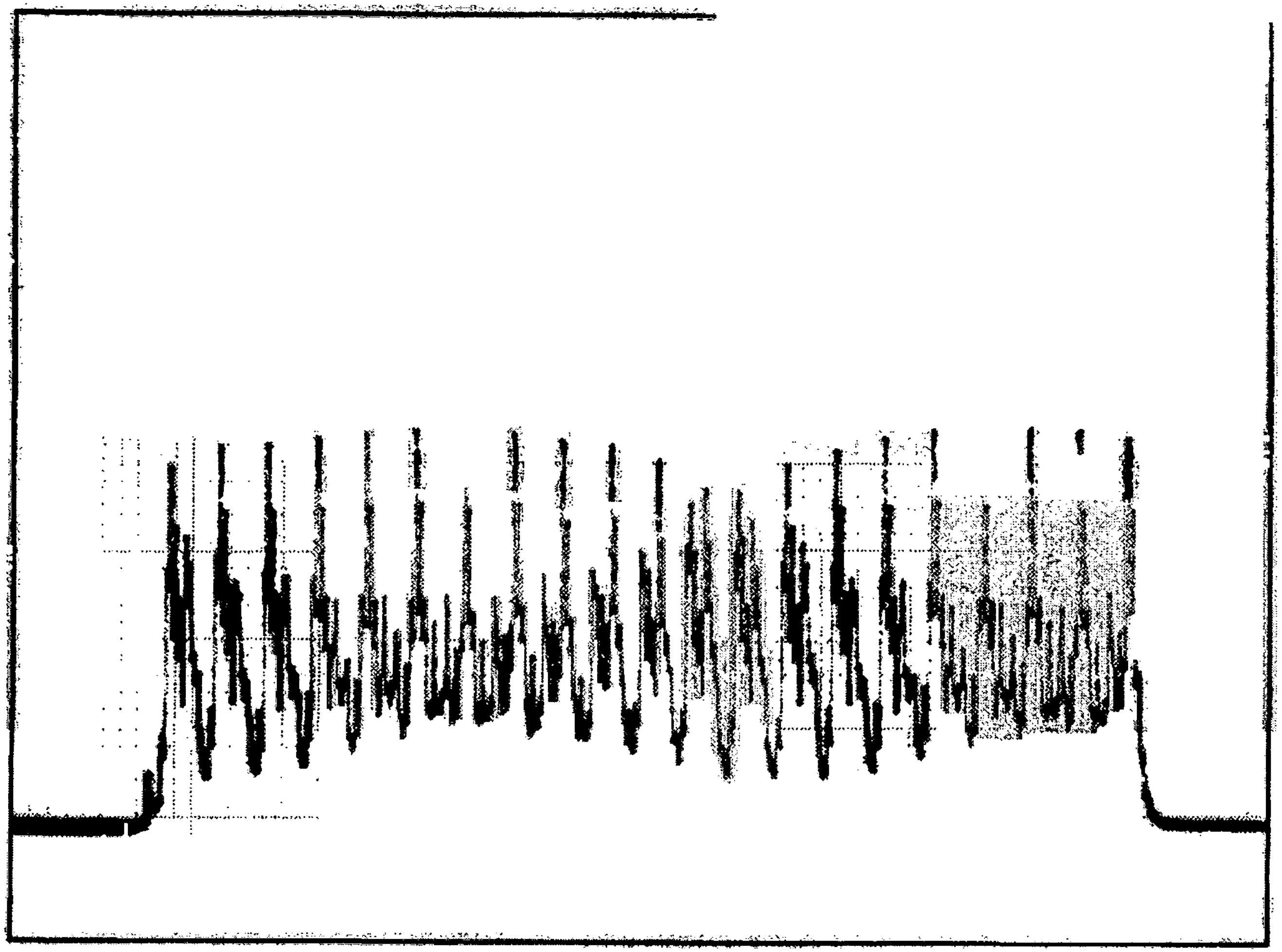


Figure 3. CVSA output when the input is an 8 millivolt peak-to-peak 200 Hz sine wave (stressed response).

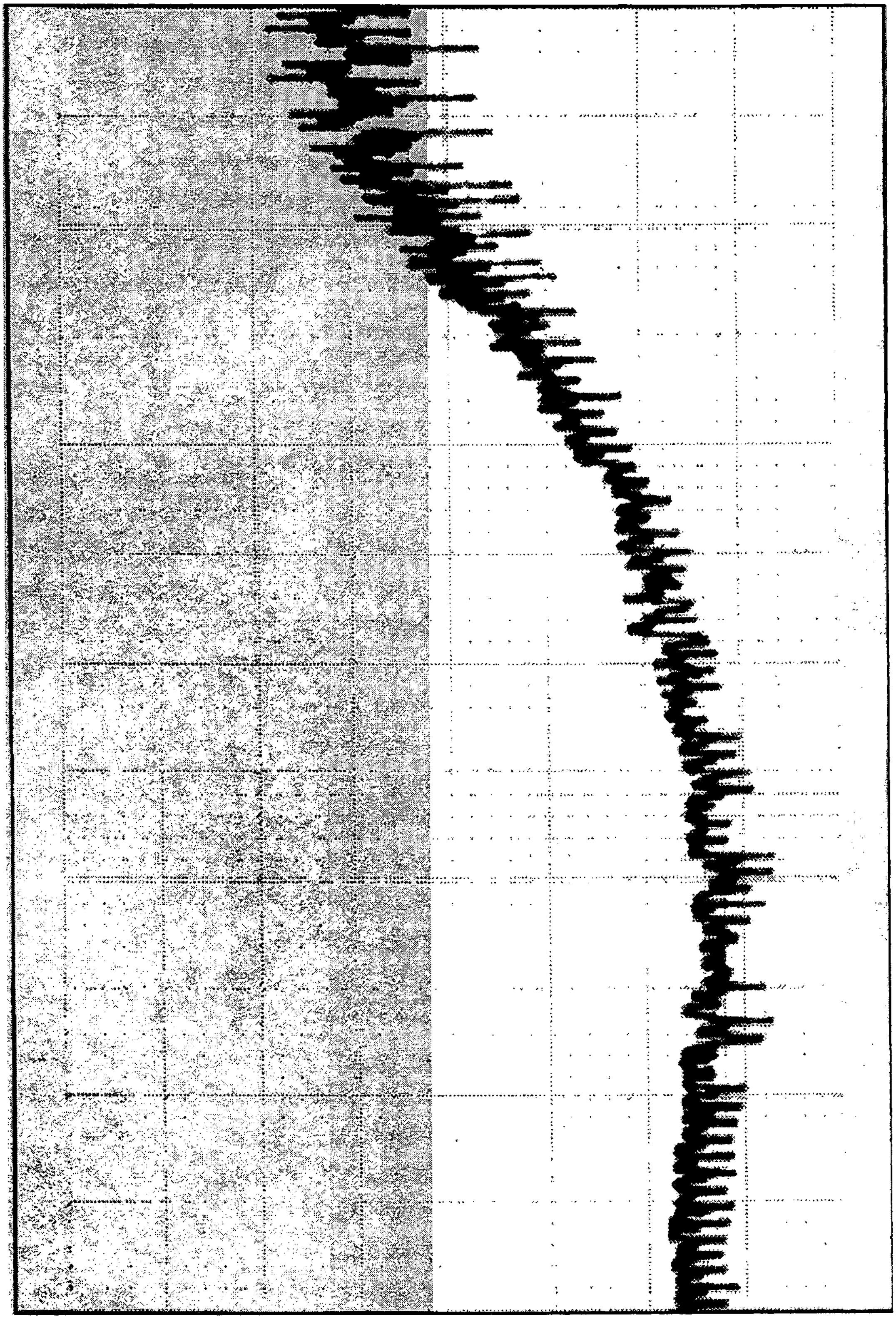


Figure 2. CVSA output when the input is an 8 millivolt peak-to-peak sine wave, swept from 500 to 2500 Hz (left to right).

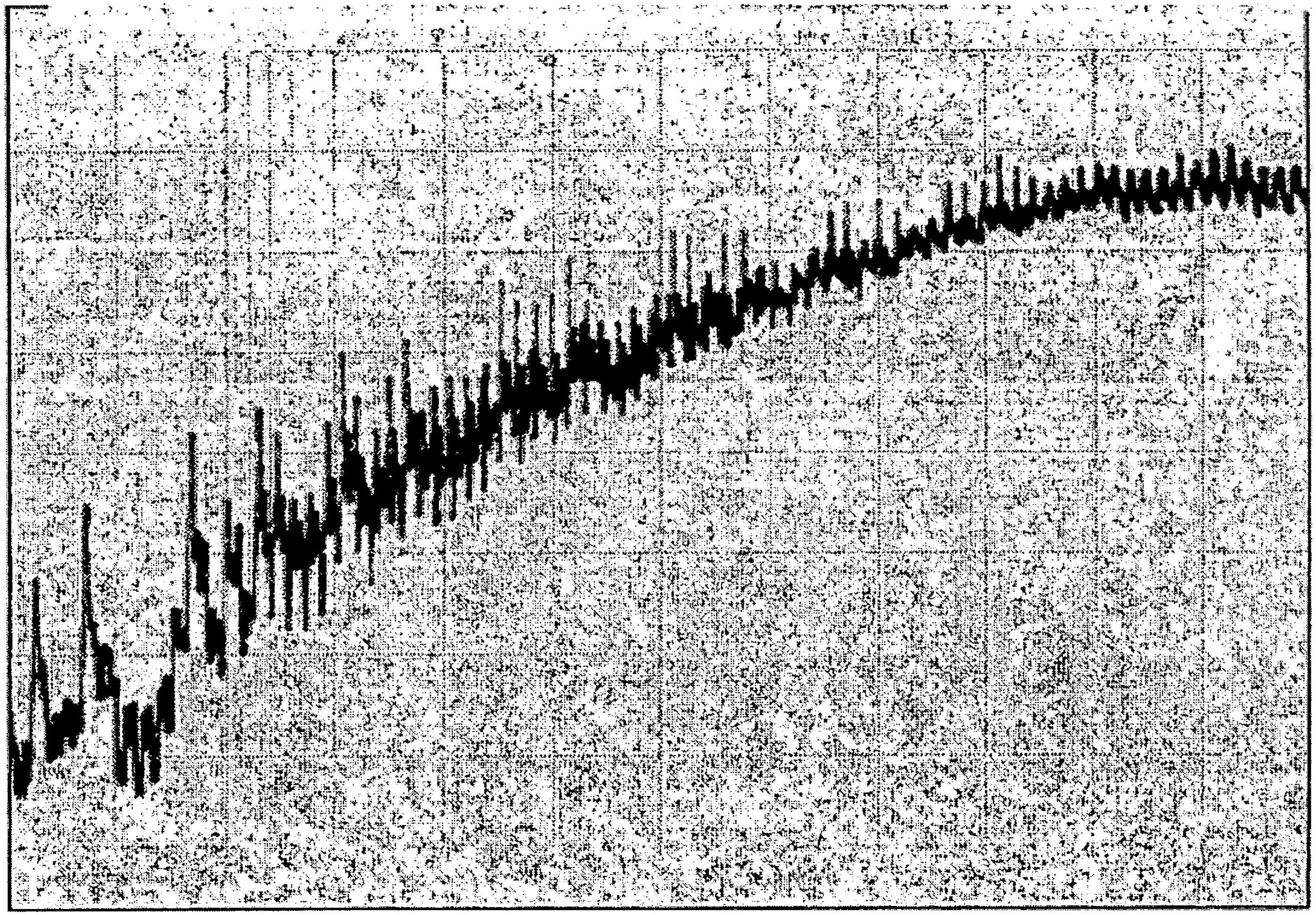


Figure 1. CVSA output when the input is an 8 millivolt peak-to-peak sine wave, swept from 100 to 1000 Hz (left to right).

to the modulation input of the CFG280 Function Generator. The starting frequency for the CFG280 Function Generator was initially adjusted with zero volts applied to the modulation input. The frequency of the signal on the output of the CFG280 Function Generator was then increased linearly by application of a positive-going linear voltage ramp at the modulation input. The CFG280 Function Generator has a modulation transfer function such that the instantaneous output frequency is a function of the instantaneous amplitude of the signal presented to the modulation input. The amplitude of that signal was adjusted so that the CFG280 provided a frequency modulated carrier signal to the CVSA, varying linearly from a starting frequency of 100 Hz to an ending frequency of 1000 Hz. The steady-state (unmodulated) amplitude of the sine wave output of the CFG280 was limited to 8 millivolts peak-to-peak except for the signal strength test, in which a 32 millivolt signal was applied. The resulting output from the CVSA was recorded on 2 inch (508 mm) heat sensitive paper normally used for recording. Frequency and time were recorded respectively on the vertical and horizontal axes of the chart paper as shown in Figures 1 through 10. The ramp test was repeated using 500 Hz and 2500 Hz, respectively, as start and stop frequencies. The response of the CVSA was also recorded using unmodulated, or continuous wave signals, and frequency modulated (FM) signals. The modulation frequency was fixed at 10 Hz to simulate microtremor with an amplitude sufficient to achieve approximately 25% modulation from the target function generator. The 25% modulation level was selected as being representative of voice microtremor magnitude. Response patterns were checked at 200Hz, 500 Hz, and 1000 Hz to assess CVSA system linearity. The sensitivity setting on the CVSA was adjusted so that the light emitting diode range indicator on the CVSA instrument front panel stayed within the normal range during signal acquisitions. Chart sizing was adjusted to provide normal pen deflections on the CVSA prior to each simulation.

Results

Laboratory test results indicate that the CVSA pen vertical position is dependent on the input frequency, within a narrow linear range (see Figures 1 and 2). When the CVSA input was an 8 millivolt sine wave, the frequency of which was increased from 100 to 1000 Hz~as illustrated in Figure 1—midscale corresponded to approximately 350 Hz. Figure 2 depicts the response relationship between frequency and pen position from 500 Hz to 2500 Hz. CVSA frequency response appears to be linear from 100 Hz to 600 Hz. Between 600 Hz and 2000 Hz, a gradual non-linear response pattern is observed in the CVSA output (Figures 1 and 2). Above 2000 Hz, the frequency response rolls off rapidly to become nearly flat.

Injecting a 200 Hz continuous sine wave signal into the microphone input of the CVSA resulted in the CVSA output shown in Figure 3. Figure 4 shows the same 200 Hz fundamental frequency input signal with 10 Hz FM modulation and approximately 25% modulation. A cyclic pattern is evidence, as predicted by the manufacturer's theory of operation. In accordance with that theory, the signal depicted in Figure 3 would correspond to a stressed (deceptive) response, while that of Figure 4 would indicate the absence of reduction of stress (truthful response). Similar response patterns are seen in Figures 5 and 6 (fundamental frequency = 500 Hz), and in Figures 7 and 8 (fundamental frequency = 1000 Hz).

Increasing the amplitude—but not the frequency—of the signal injected into the CVSA microphone input (i.e., an increase in signal strength) resulted in a decrease in the amplitude modulated baseline component of the CVSA output, as shown in Figures 9 and 10. The signal riding on the detected FM may be front end or discriminator noise, which is reduced in amplitude as the input signal amplitude is increased (i.e., FM quieting increases as the signal to noise ratio gets larger). In the CVSA the undetected AM component (noise) is largely a function of variations in carrier level (fundamental frequency amplitude). The FM detector responded primarily to changes in the carrier frequency (microtremor), with an output voltage change proportional to the change in frequency (Figures 4, 6, and 8).

microtremor, indirectly assessed by analysis of changes in voice fundamental frequency, is purported to be inversely related to stress (Brenner, et al., 1979; Inbar & Eden, 1976; Smith, 1977; VanDercar, et al., 1980). It is argued that as stress increases, the amplitude of the microtremor decreases. Support for the laryngeal microtremor hypothesis is inconsistent (Inbar & Eden, 1976; Shipp & Izdebski, 1981). Shipp and Izdebski (1981) found no evidence to support the laryngeal microtremor hypothesis. These investigators examined electromyographic (EMG) activity directly from the laryngeal muscles (cricothyroid and posterior cricoarytenoid) during conversational speech and sustained phonation. They contend that EMG activity changed so rapidly over time during normal speech that no Fourier analysis could be calculated at the selected sampling rate. These signals were compared to normal microtremor of 9 Hz sampled from the biceps. They concluded that their findings cast doubt on the assumptions made by manufacturers of (voice) stress analysis instruments. Conversely, Inbar and Eden (1976), using similar procedures, found that EMG recordings were correlated with frequency changes in the voice spectrum, suggesting the existence of voice microtremor.

Voice stress research using instrumentation other than off-the-shelf voice stress analyzers has focused on discrete measures within the response as indicators of deception. Motley (1974) reported that response duration was the only reliable index of deception. Other investigators have shown that stress is related to a specific change in the fundamental frequency of the speaker's voice (Tolkmitt & Scherer, 1986; Streeter, et al, 1977). Cestaro and Dollins (1994) calculated spectrum and time domain analyses of voice responses recorded during 28 peak of tension (POT) psychophysiological detection of deception (FDD) examinations. They found no single measure of the voice response that could serve as a reliable indicator of deception. While a systematic relationship was found among some combinations of speech parameters and stress, the relationship was not consistent over time and between subjects. No systematic relationship was found between voice spectra and stress. Others have reported that speaker stress could not be related to the results of voice response spectral analysis (Zalewski, Majewski, & Hollien, 1975).

This study was designed to evaluate the published theory of operation of a second generation voice analyzer, the CVSA. According to the manufacturer, the CVSA detects stress related changes in the voice (laryngeal) microtremor (NITV, 1994). The experiment was designed to determine whether the CVSA instrument detects microtremor in the fundamental frequency of presented signals. The manufacturer claims that changes in the fundamental frequency of a signal presented at the input to the instrument are displayed as meaningful changes in the chart tracings; tracings with a constant, or nearly constant, amplitude (i.e., containing little or no microtremor) are indicative of a stressed response (NITV, 1994). Conversely, tracings showing a cyclic or peaked pattern are claimed to be the result of microtremors in the response, and are indicative of a response containing little or no stress. Laboratory function generators were used to present simulated stressed and unstressed voice responses to the input of the CVSA. Constant amplitude unmodulated signals were used to represent a stressed voice response containing no microtremor. Unstressed responses were simulated by frequency modulating the function generator at a 10 Hz rate. The resultant CVSA output was examined at various fundamental frequencies.

Method

Apparatus

A Tektronix (Beaverton, OR) Model CFG280 Function Generator was used to present a constant amplitude sine wave signal to the microphone input of the CVSA. A Tektronix Model CFG250 Function Generator was used to modulate the frequency of the CFG280 signal and simulate speech microtremor. Frequency was verified with a Tektronix CDC250 Universal Counter. A Tektronix Model 2247 Oscilloscope was used to monitor the amplitude and frequency shift of the signals from die CFG280 Function Generator.

Procedures

The CFG250 Function Generator was used to present a linear ramp (sawtooth) input

A Test of the Computer Voice Stress Analyzer (CVSA) Theory of Operation

Victor L. Cestaro

Abstract

This study was designed to test the underlying electronic theory of operation of the Computer Voice Stress Analyzer (CVSA). During this experiment the CVSA input/output was evaluated using simulation signals from laboratory test generators. The laboratory simulations established that the CVSA performs electrically according to the manufacturer's theory of operation. These results indicate there may be a systematic and predictable relationship between displayed voice patterns and changes in the speech envelope related to human physiology.

Voice analysis is the decomposition of a human voice into objectively measurable characteristics. It has been proposed that voice analysis can determine the amount of stress (voice stress analysis) that the speaker is experiencing (Brenner, Branscomb & Schwartz, 1979; Inbar & Eden, 1976). Further, it has been suggested that voice stress is linked to deception (Motley, 1974; O'Hair & Cody, 1987; Streeter, Krauss, Geller, Olson, & Apple, 1977). It should be noted that instruments designed to detect deception, such as the polygraph instrument, do not detect deception per se, but rather detect physiological activity related to the stress experienced by subjects during the act of deception.

The Computer Voice Stress Analyzer (CVSA) manufactured by the National Institute of Truth Verification (NITV, West Palm Beach, FL) is the latest in a series of instruments purported to detect deception in voice responses. Previous equipment, such as the Psychological Stress Evaluator (PSE), consisted mainly of a simple resistor-capacitor low pass filter circuit, and required responses to be recorded on audio tape and subsequently analyzed at reduced tape speed (VanDercar, Greaner, Hibler, Spielberger, & Block, 1980). Unlike the PSE, the CVSA analyzes and displays responses in real time, purportedly using state of the art computer technology. Responses do not have to be pre-recorded and then played back at the 1/4 to 1/8 speed required by the PSE.

The underlying theory of operation for the PSE and the CVSA is that the instruments detect physiological microtremor associated with muscles in the voice mechanism. Physiological tremor is described as a low amplitude oscillation of the reflex mechanism that controls the length and tension of a stretched muscle, and has a frequency between 8 and 12 hertz (Hz)(Lippold, 1971). According to Lippold, tremor is believed to be a function of the signals to and from motor neurons; it is analogous to a self-adjusting, closed-loop servo system. That is, the observed tremor is like the "hunting" behavior of mechanical servomechanisms. Stretch sensors in the muscle tissue signal the amount of stretching and transmit this information to the associated motor neuron in the spinal cord. This information is processed and the efferent motor neuron fiber is activated to increase or decrease the stretch of the muscle tissue. The finite delays in signal transmissions to and from the target muscle account for the low frequency oscillation, and hence, the hunting behavior.

Voice stress analyzers purportedly detect physiological microtremor in speech (oscillations of 8 to 12 Hz in muscle tissue), and convert those components to a graphical representation of stress experienced by the subject (Brenner, Branscomb, & Schwartz, 1979). Nerve fibers carried in the trunk of the vagus nerve innervate the laryngeal muscles, including the cricothyroid muscle (Kahane, 1986). Increases in voice frequency are accomplished by lengthening the vocal folds through activity of the cricothyroid muscle, while decreases are a result of relaxation and shortening of the vocal folds by the thyroarytenoid (Gray, 1977, p. 963). Laryngeal

Streeter, L.A., Krauss, R.M., Geller, V., Olson, C., & Apple, W. (1977). Pitch changes during attempted deception. *Journal of Personality and Social Psychology*, 35. 345-350.

Tolkmitt, F.J., & Scherer, K.R. (1986). Effect of experimentally induced stress on vocal parameters *Journal of Experimental Psychology*, J_2, 302-313.

response intensity, response duration, and the FM energy component may prove to be a reliable additional polygraph channel. Speech formant structures and a more stringent analysis of spectrum data should be examined in further studies, and added to the final equation. Computer programs employing neural networks, fuzzy logic, or other "smart" procedures may, in the future, identify response characteristics within a polygraph session and adjust weights accordingly to provide increased levels of confidence in that channel's decision output. However, the results of this research, and of the reviewed studies, suggest that voice stress analysis within the context of a standard PDD examination is not yet a reliable and valid discriminator of truth and deception.

Acknowledgements

Requests for reprints should be sent to Dr. Victor L. Cestaro, Building 3195, Department of Defense Polygraph Institute, Fort McClellan, Alabama 36205-5114. The authors wish to express special thanks to the volunteers who participated in this study: Charlene L. Stephens, Master Sergeant Randall S. Reynolds, Dr. Donald J. Pettit, and Larry R. Broadwell for their assistance throughout data acquisition and processing. This research was supported by DODPI94-P-0024 project funds from the Department of Defense Polygraph Institute. The results of this study were previously published as a U.S. Government Technical Report (Cestaro & Dollins 1994). The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government

References

Barland, G.H. (1978). Use of voice change in the detection of deception. *Polygraph*, 7, 129-140.

Brenner, M., Branscomb, H.H., & Schwartz, G. (1979) Psychological stress evaluator - two tests of a vocal measure. *Psychophysiology*, 16, 351-357.

Cestaro, V.L., & Dollins, A.B. (1994). An Analysis of Voice Responses for ihe Detection of Deception (Report No. DODPI-R-OOOI). Fort McClellan, AL: Department of Defense Polygraph Institute.

Dollins, A.B., Cestaro, V.L., & Perth, D. (1994). Efficacy of Repeated Psychophysiological Detection of Deception Testing (Report No. DODPI-R-0013). Fort McClellan, AL: Department of Defense Polygraph Institute.

Fay, P.J., & Middleton, W.C. (1941). The ability to judge truth-telling or lying, from the voice as transmitted over a public address system. *The Journal of General Psychology*, 24, 211-215

Horvath, F. (1982). Detecting deception: The promise and the reality of voice stress analysis. *Journal of Forensic Science*, 27, 340-351 Reprinted in *Polygraph*, JJ., 304-318.

Lieberman, P. (1961). Perturbations in vocal pitch. *The Journal of the Acoustical Society of America*, 33, 597-603.

Lieberman, P., & Michaels, S.B. (1962). Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *The Journal of the Acoustical Society of America*, 34, 922-927.

Motley, M.T. (1974). Acoustic correlates of lies. *Western Speech*, 38, 81-87.

O'Toole, G. (1975). *The Assassination Tapes*. New York. Penthouse Press, Ltd.

Spectrum Data Analysis

Average magnitudes within 200 Hz bins (partitions) across the maximum allowable passband (5000 Hz) for the selected sampling rate (10 Khz) were calculated. A rank order assignment of bin magnitudes, with 1 representing the highest magnitude bin and 25 representing the lowest magnitude bin in serial order from 1 Hz to 5000 Hz, was made to generate a profile of responses for each question and subject, within a test. Since this was a relative measure, overall differences in response voice amplitude were not expected to be confounding factors.

Profiles for deceptive and non-deceptive responses were compared for congruence within each subject's data set. The dependent measure was the serial alignment (pattern match) of the 25 ranked bin values for each question with the mean ranking of the five question set. Serial alignment was assessed by non-parametric correlation (Spearman rho) The greatest pattern mismatch was expected to be associated with the question causing the most stress to be subject. A correlation of -1.0 indicates a severer misalignment of patterns, and a correlation of 1.0 is indicative of an exact pattern match. Although correlations in the direction of misalignment were seen in some cases, no systematic mismatch was found for deceptive responses to the target question.

Discussion

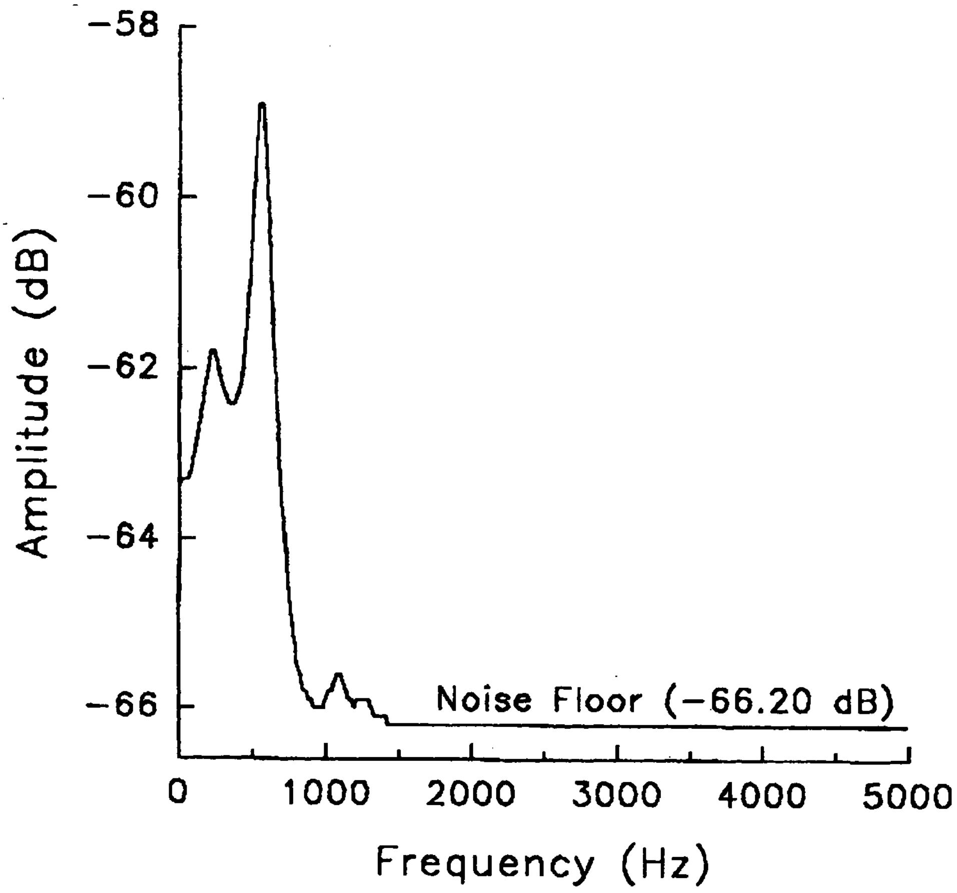
Results indicated that no single human voice measure, as collected and evaluated in this study, reliably discriminated between truthful and deceptive responses. The measures examined include: dominant (fundamental) pitch frequency, voice response energy, response duration, and the magnitude and frequency of pitch changes. Within the groups sampled, the FM component had a range of 0.6 to 28.8 Hz. However, other investigators have reported that the FM component studied by Psychological Stress Evaluators (PSE) has a range of 8 to 14 Hz (Brenner, Branscomb, & Schwartz, 1979). It is, thus, not clear whether this FM component is equivalent to the PSE or is a measure of some other component.

Although other investigators have reported that a short duration response was a reliable indicator of deception (e.g , Motley, 1974), the results of the present study indicate that duration is an unreliable index of deception. Response duration may be susceptible to cognitive countermeasures (e.g., intentional changes in response duration). Changes in voice intensity (speech amplitude) were not indicative of deceptive responses and may also be susceptible to countermeasures. Various pitch parameters, however, are associated with parasympathetic nervous system activity (the vagus nerve innervates the laryngeal muscles controlling certain aspects of speech), and are not under voluntary control. Streeter, et al. (1977) found that the FO of subject responses was higher during deceptive than non-deceptive responses. That relationship was not found in this study. However, instantaneous changes in the fundamental pitch frequency, and the magnitude of those changes may be related to emotional arousal or stress. The FM energy component, derived from the instantaneous change measure and magnitude, may serve as a more reliable indicator of truth or deception than any single voice measure.

Lieberman and Michaels (1962) reported that the ability of observers to correctly identify emotional states of subjects dropped significantly when all pitch information was removed from subjects' recorded responses. In the present study no significant relationship was found between the FM energy component, derived from pitch, and deceptive responses. However, a higher correct decision rate was found when the FM energy component was compared to any of the single measures investigated. Since the verbal responses were collected during a peak-of-tension polygraph examination, and only a single voice response was recorded immediately after each question, there may not have been sufficient time for a stress response to appear in the recorded waveform. Further investigations might employ a restructured question format with more than one response after each question, or instructions to subjects to delay their verbal responses This may increase the likelihood that a delayed stress related response will be captured.

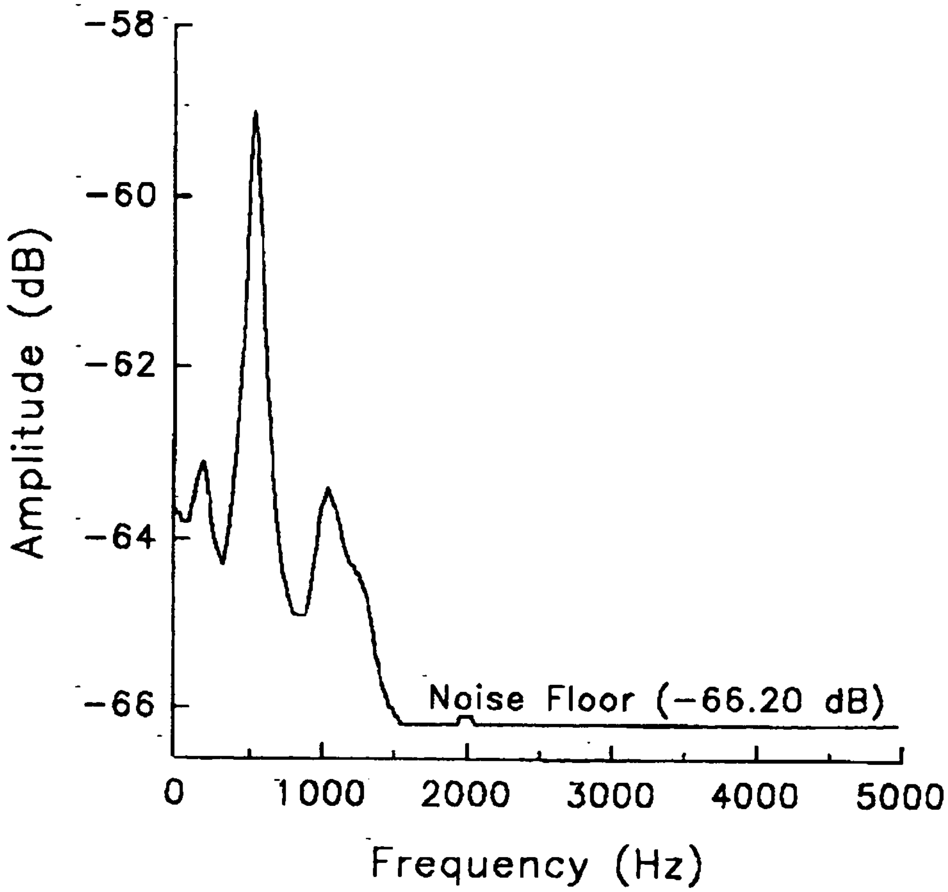
A weighted combination of mean

Figure 8



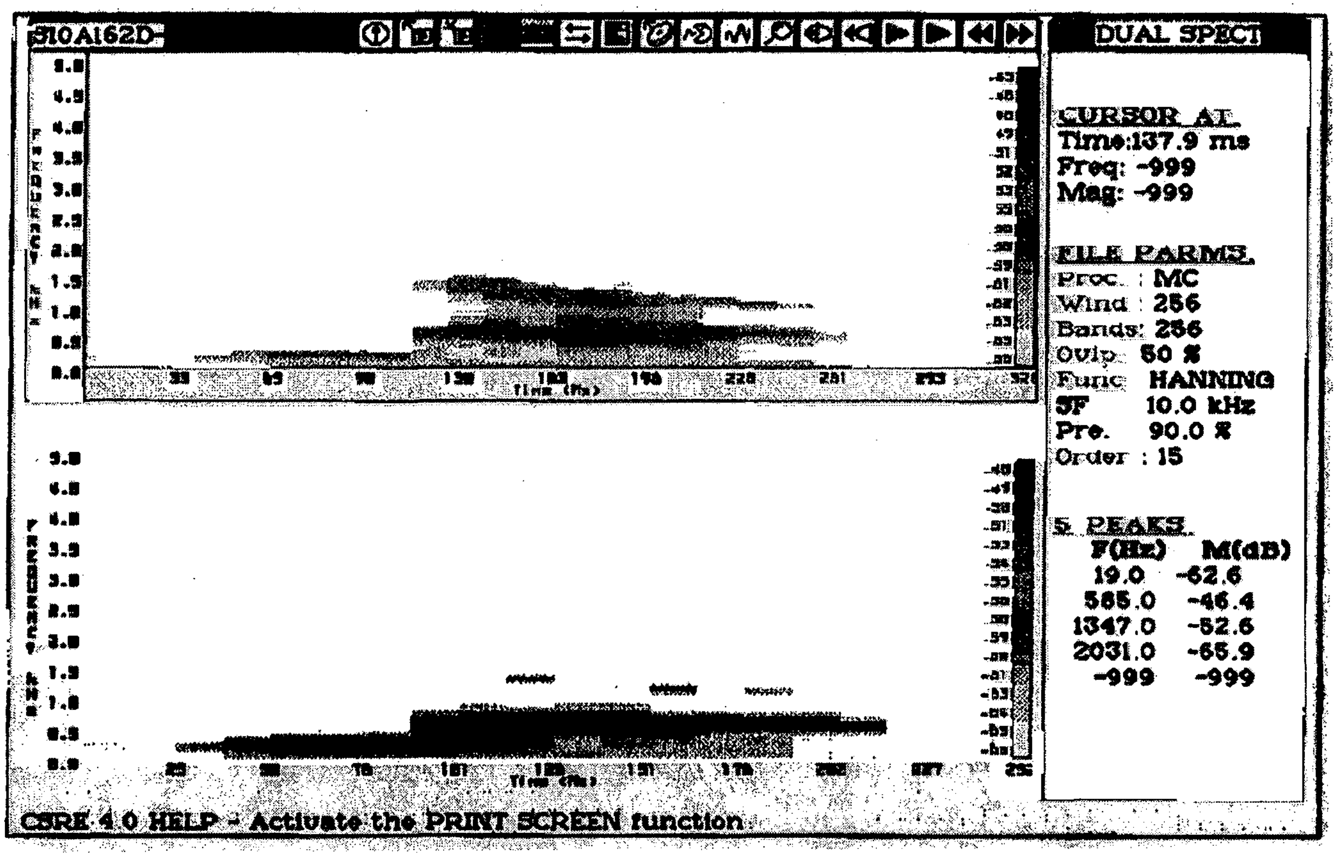
Simple spectrograph showing the same (Figure 7) deceptive subject's response to a target number question. Note the absence of energy about 1100 Hz.

Figure 7



Simple spectrograph showing a deceptive subject's response to a non-target number question. Note the energy peak at 1100 Hz.

Figure 6



CURSOR AT
Time:137.9 ms
Freq: -999
Mag: -999

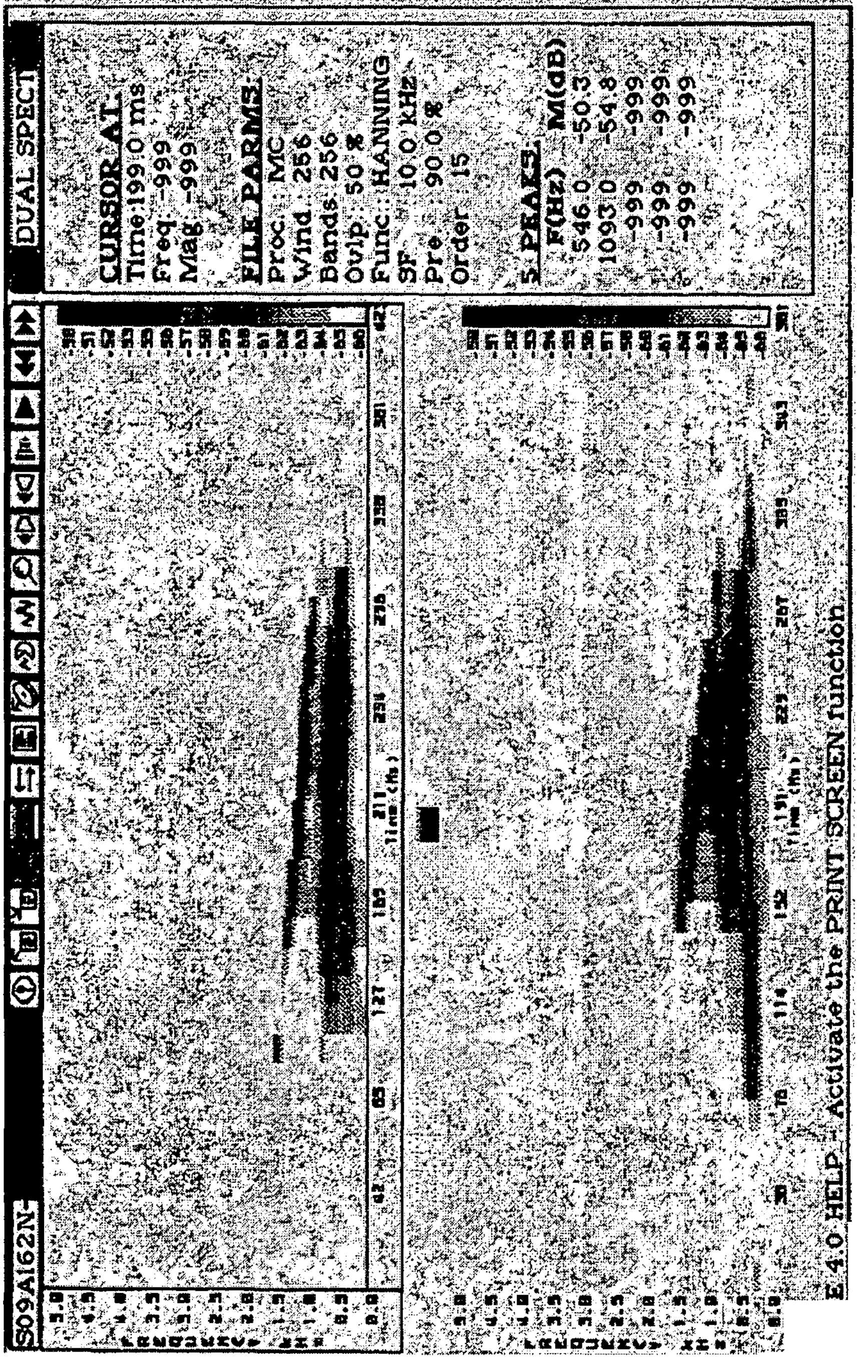
FILE PARMS
Proc. : MC
Wind : 256
Bands: 256
Ovlp. 50 %
Func HANNING
SF 10.0 kHz
Pre. 90.0 %
Order : 15

5 PEAKS

F(Hz)	M(dB)
19.0	-62.6
585.0	-46.4
1347.0	-52.6
2031.0	-65.9
-999	-999

Complex spectrograph showing a deceptive subject's responses to a non-target (upper panel) and a target number question (lower panel).

Figure 5



CURSOR AT:
Time:199.0 ms
Freq: -999
Mag: -999

FILE PARMS:
Proc: MC
Wind: 256
Bands: 256
Ovlp: 50 %
Func: HANNING
SF: 10.0 kHz
Pre: 90.0 %
Order: 15

5 PEAKS:
F(Hz) M(dB)
546.0 -50.3
1093.0 -54.8
-999 -999
-999 -999
-999 -999

E 4.0 HELP - Activate the PRINT SCREEN function.

Complex spectrograph showing a non-deceptive subject's responses to a non-target (upper panel) and target number question (lower panel).

Pitch Data Analysis

A certified forensic psychophysiologist at the DoD Polygraph Institute independently examined subjects' physiological data to determine which number was circled by each subject. His determinations were based on chart tracings of two pneumo channels, the cardio channel, and the GSR channel. Where no determination could be made by the examiner, the data were dropped from the analysis, leaving 50 tests out of a possible 84 for an analysis of agreement rates. The frequency of concurrent determinations (i.e, a numbers match) made by the examiner and the FM energy index was significantly different from chance expectation (Z = 4.0, p < .01, two-tailed). In other words, both the examiner and the energy index identified some response to a particular number, whether or not it was the number circled by the subject during the anagram task. No attempts were made to determine whether a subject's responses were evaluated as DI (deception indicated) or NDI (no deception indicated) during this comparison.

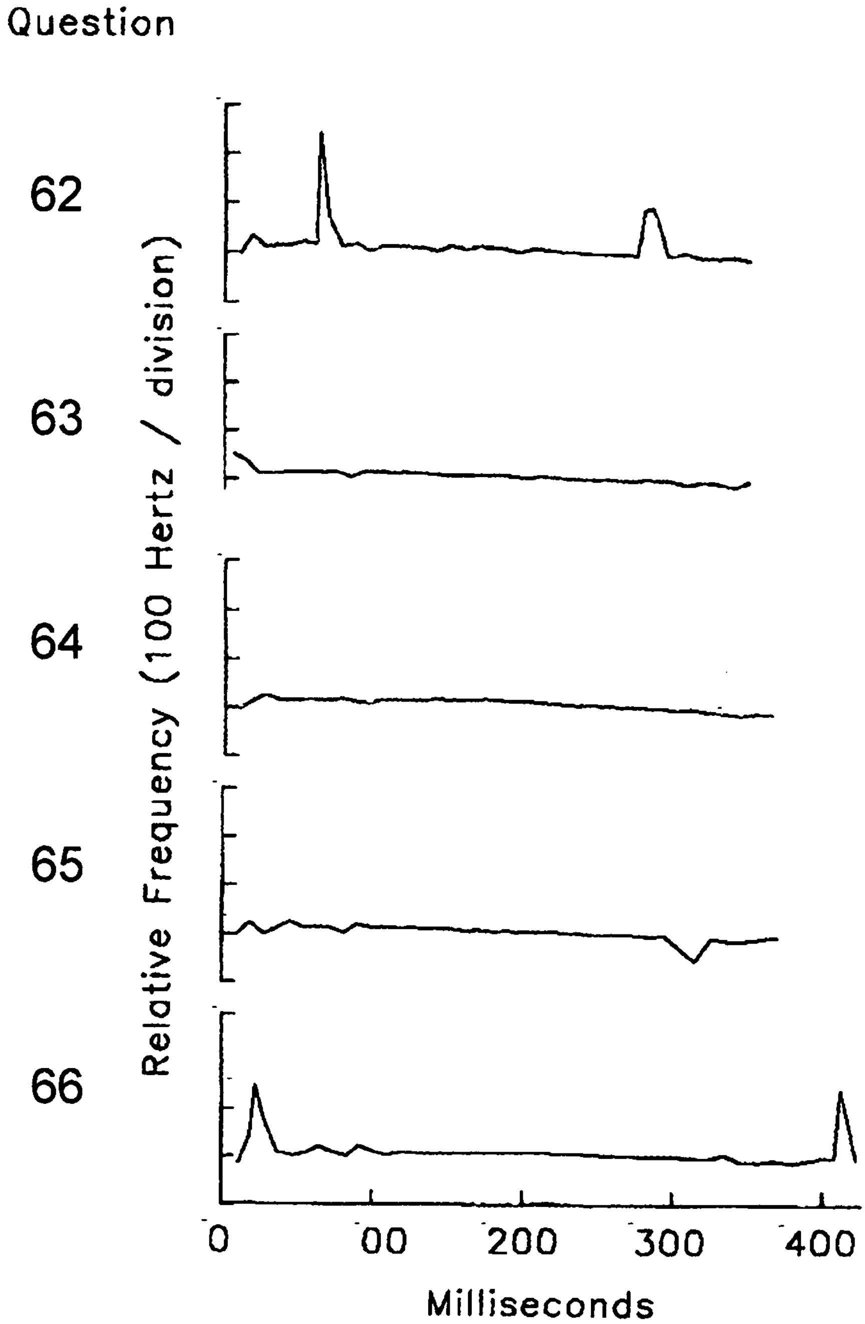
Examination of the above "correct number" decisions showed that, based on subject programming of DI (target number denied by subject) and NDI (subject's target number omitted from test), the examiner had 79% correct DI decisions versus 37% correct DI decisions based on pitch/energy ranking (Z - 3.46, p. < .05, two-tailed). This result indicated that there was a significant difference between the frequency of correct target number determinations made by the examiner and by the pitch/energy ranking algorithm. Further analysis indicated that the frequency of correct number determinations using the pitch/energy ranking algorithm was not significantly greater than chance. However, the examiner had a 35% false positive rate versus a 29% false positive rate using the pitch/energy ranking algorithm (Z = .375, p. > .05, two-tailed), demonstrating that there were no significant differences between the false positive rates of the two methods. There were only two cases where both the examiner and the pitch/energy ranking method concurred on a false positive decision. Three separate GROUPS (2) x TEST (3) x QUESTION (5) repeated measures analyses of variance revealed no significant differences for measures of dominant frequency, energy, or duration.

Visual Analysis of Spectrograms

The spectrographs were printed and subsequently analyzed by overlaying and visually inspecting the degree of spectrograph match-mismatch. Figure 5 shows the spectrographs for a subject programmed non-deceptive, with the upper spectrograph depicting the non-target number response, and the lower showing the response to the target number question. Figure 6 shows the response patterns of a subject programmed to be deceptive.

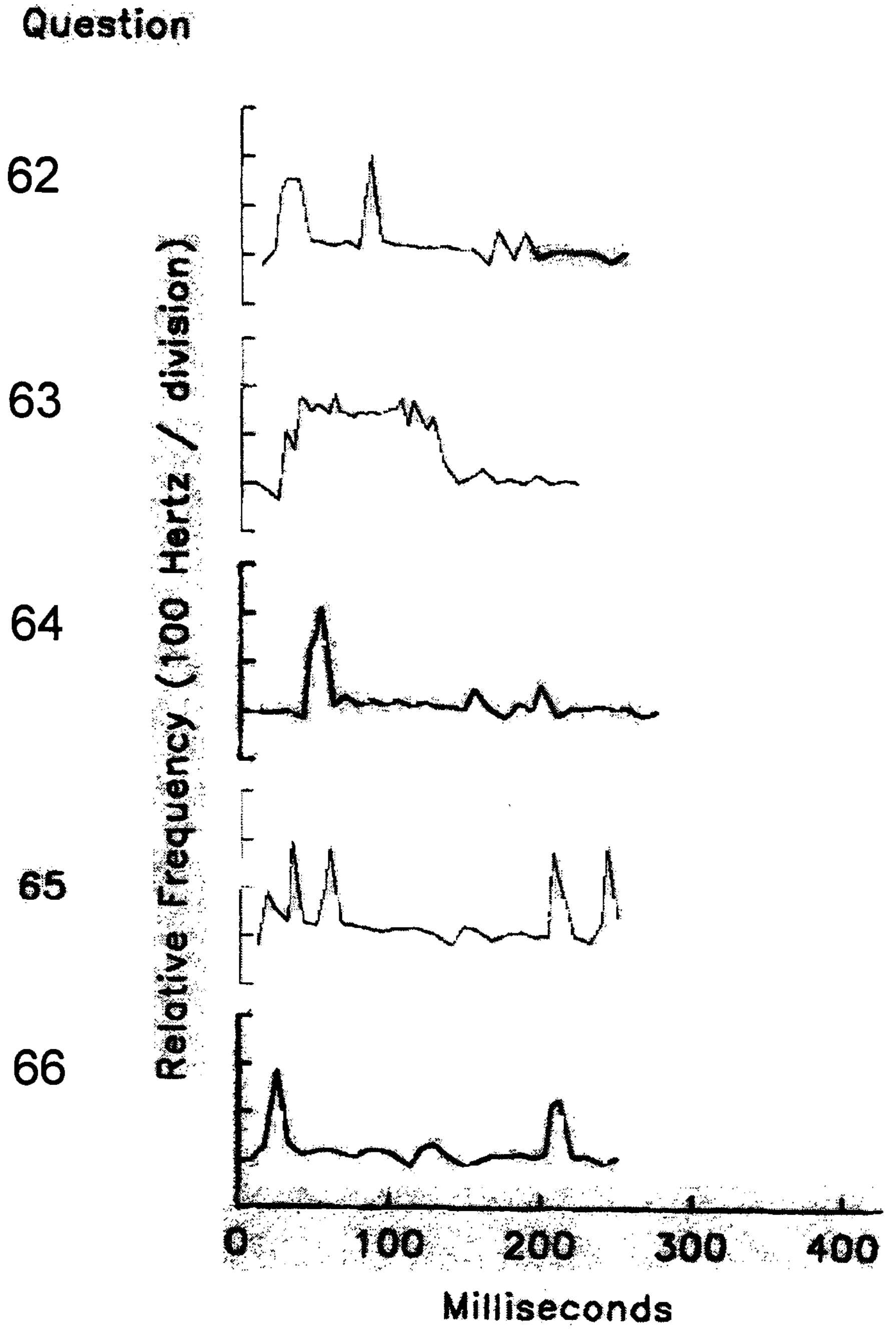
Since visual inspection was determined to be too inaccurate for objective analyses, the data were collapsed across time to produce a standard amplitude x frequency spectrograph Figures 7 and 8 are amplitude x frequency spectrographs of the data displayed in the complex spectrograph (Figure 6). The amplitude x frequency information was then divided into a series of partitions for statistical and pattern analyses.

Figure 4



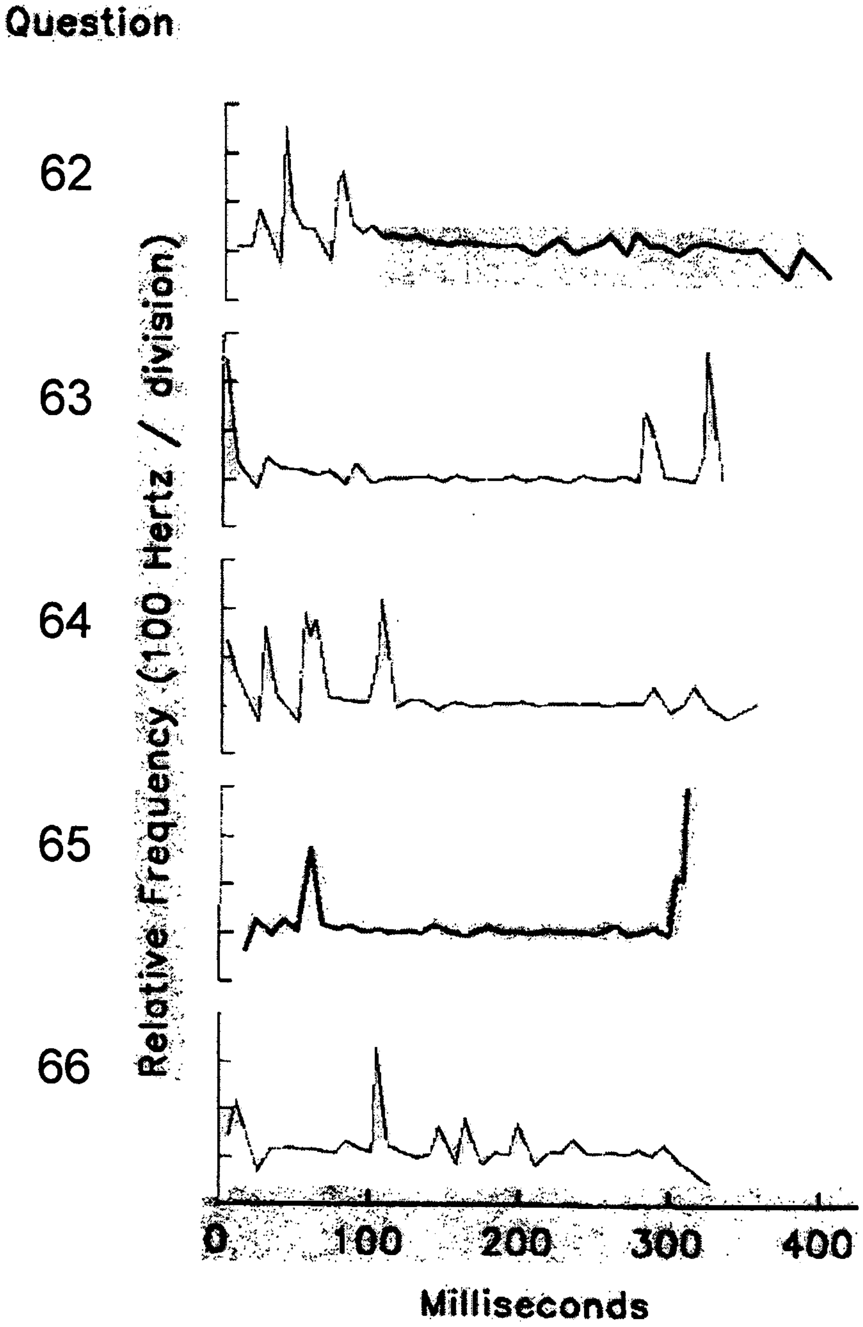
Pitch contours of a second non-deceptive subject's responses to five questions, showing an absence of pitch variations in the response to the target number question (64).

Figure 3



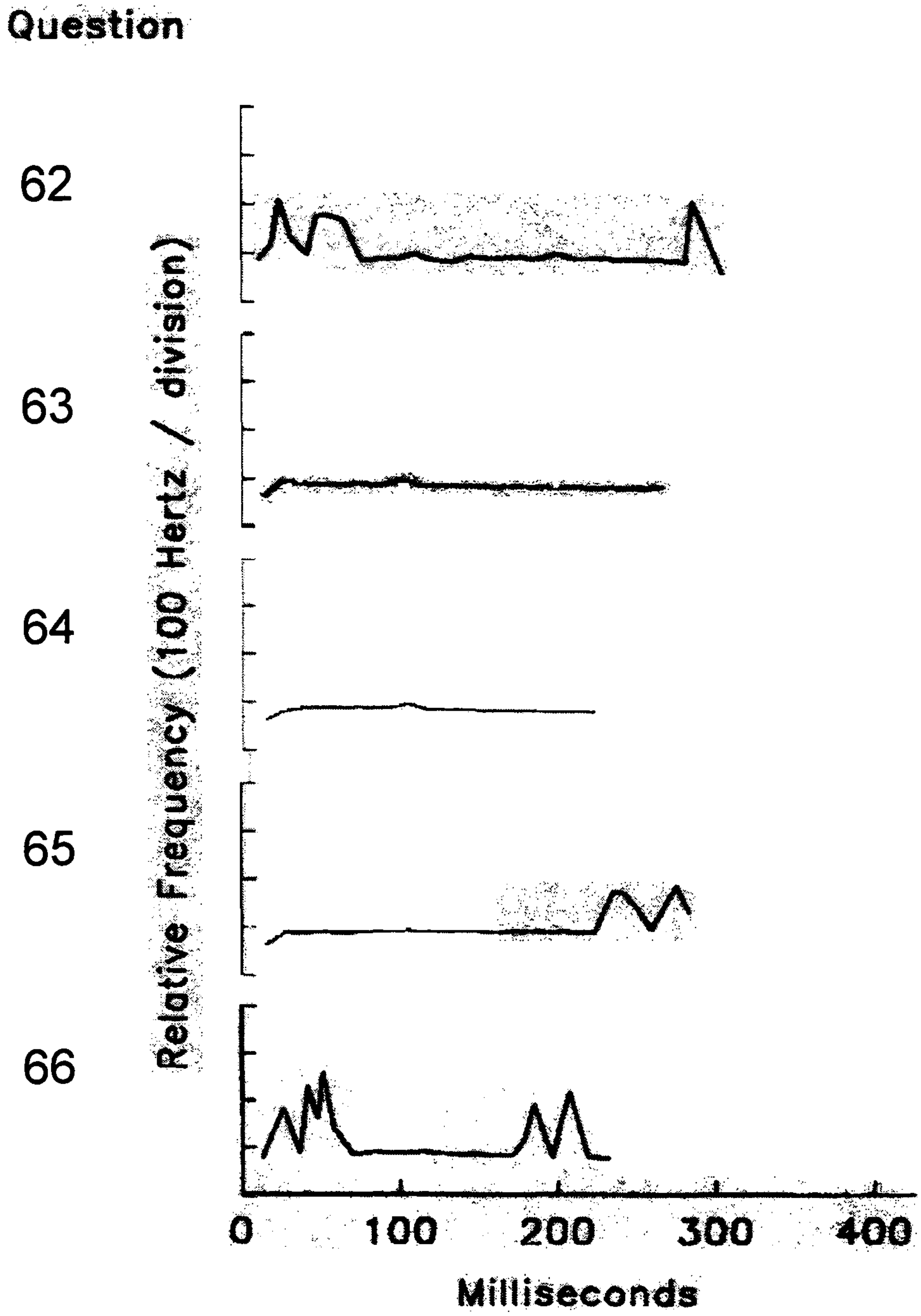
Pitch contours of a second deceptive subject's responses to five questions, showing pitch variations during all responses.

Figure 2



Pitch contours of a non-deceptive subject's responses to five questions, showing large pitch variations during all responses.

Figure 1



Pitch contours of a deceptive subject's responses to five questions, showing an absence of pitch variations in the response to the target number question (64).

with an examiner's visually-based decisions.

Spectrum Analysis Data Reduction

The CSRE software was designed to perform spectral analyses of speech, employing Fast Fourier Transforms (FFTs), Modified Covariance (MC), and Autocorrelation (AC) techniques The resulting spectral pattern can be displayed on a computer screen using a magnitude (in dB) x frequency (in Hz) x time (in ms) scale. Spectrum data files were saved on computer disk for additional off-line processing.

It was determined during trial analyses that the Modified Covariance technique was the optimum method of spectrum decomposition for short duration responses. This method is also recommended by the software manufacturer. Signal pre-emphasis was set at 90% to compensate for approximately 6 dB per octave roll-off of voiced speech, largely due to radiation at the lips.

Results

Visual Analysis of Pitch Waveforms

Graphics software was used to examine the continuous pitch contours of the five responses within a test (Figures 1, 2, and 3).

It can be seen that the pitch waveform of a programmed "deceptive" subject's response (Figure 1) is represented by a relatively straight line during the target number question response (middle waveform), with little change in the dominant pitch frequency. However, the responses to questions before and after the target show obvious deviations (FM component) from the dominant frequency, especially during responses to the questions concerning the numbers 62, 65, and 66. This was not the case for responses from a programmed "non-deceptive" subject (Figure 2). All five waveforms recorded from the subject contained obvious deviations from the dominant pitch frequency. However, in many cases, subjects programmed "deceptive" showed the same pattern of responses as a programmed "non-deceptive" subject (Figure 3), and in others, the opposite pattern as seen (Figure 4).

sensitivity adjustments were made.

The following series of statements were made and questions asked, via computer recorded voice, during a single chart:

X The test is about to begin.
01 Did you complete an anagram for the number 60?
02 Did you complete an anagram for the number 61?
03 Did you complete an anagram for the number 62?
04 Did you complete an anagram for the number 63?
05 Did you complete an anagram for the number 64?
06 Did you complete an anagram for the number 65?
07 Did you complete an anagram for the number 66?
XX The test is now complete, please continue to sit still white I turn the instrument off.

Before the examination began, the examiner reminded each subject that the correct response to each question was displayed on the wall directly in front of the subject. If the examiner judged that the physiological signals recorded on the polygraph chart contained artifacts, the previous question was repeated. The examiner played the message, "Please remain still" if he judged that the examinee was producing unnecessary and/or excessive movements. When a question series was completed, the pressure in the occlusive cuff was vented and the subject was instructed to "please relax while I prepare for the next test." If subjects appeared to be sleepy, they were also reminded of the importance of the study and encouraged to remain alert. The next FDD test was begun approximately three minutes later. The occlusive cuff was inflated prior to beginning the next test, as described above. This procedure was repeated until six tests were completed, after which the sensors were removed. The subjects were then asked to read and sign a debriefing form, reminded to return the following week, and escorted out of the building.

Participants returning for a second test were escorted to a briefing room where they were reminded of the number circled during the previous session and asked to conceal the second card, indicating the number circled, in a pocket. They were reminded not to reveal what the number was to the examiner, then escorted to the examination room. The examiner again reviewed the biographical/medical questionnaire from their previous session to ensure that no changes had occurred. Six additional PDD tests were completed, as described above. When the examination was complete, participants were thanked for their cooperation, asked to read and sign a second debriefing form, and escorted out of the building

Pitch Data Reduction

Digitized voice responses were processed with CSRE's software comb filter to extract pitch from the raw waveform data. The data acquisition sampling rate was set to 10 Khz. The low-pass filter cutoff frequency was set to 800 Hz prior to smoothing and comb filtering. Extracted pitch waveforms were saved for off-line processing.

Response duration was the unit of time used to convert the number of peaks per response to frequency. The number of peaks per response was determined using a software peak/trough detection algorithm, therefore providing a means to detect deviations from the dominant (fundamental) pitch frequency. This provided a measure of the mean frequency modulated (FM) component of the voice waveform.

The peak excursions (deviation magnitudes) from the dominant baseline frequency were also measured. The mean peak deviation, in cycles, from the dominant pitch frequency was divided by the dominant pitch frequency to determine the modulation index of each sample (eg., a deviation of 40 Hz from a 400 Hz dominant pitch frequency = 0.10, or 10% modulation index) This result was then multiplied by the FM component to provide an index of FM energy for each response, normalized over a one second period. Simply stated, the FM component provided a measure of the rate of shift in the dominant pitch component and the modulation index provided a measure of the magnitude of that shift. The index of FM energy was used to rank order the five responses within each test for comparison each group participate in every fourth examination. That is, no more than three control or treatment group participants were tested consecutively. Twenty-two subjects were assigned to each group. Each volunteer participated in two examination sessions. The two sessions were separated by at least five working days. Subjects completed six PDD tests during each examination session. Only the responses to the numbers 62 to 66 of the first three FDD tests of the first examination session were used for voice analysis in this study. The first two responses (to numbers 60 and 61) were excluded from all analyses to avoid inclusion of possible orienting responses in subjects' data.

Upon arrival at the DoD Polygraph Institute (Fort McClellan, AL), each participate was escorted by one of the investigators to a secluded briefing room and asked to read a brief description of the research project. Individuals indicating that they would participate were asked to read and sign an informed consent affidavit. Any questions were then answered. A brief biographical/medical questionnaire was then completed, to ensure that the participant was in good health and not currently taking medication which could interfere with the PDD examination results.

The participant was required to complete a number search task, which was referred to as an anagram task. During this task, the participant circled six sequences of a two-digit number which was repeated five consecutive times (in any direction) in a 20 x 30 matrix of two digit numbers. The matrix consisted of numbers between 60 and 69 for the programmed guilty subjects — who circled the number 64, and 80 to 89 for the programmed innocent subjects ~ who circled the number 84. When the anagram task was complete, the participant was asked to write his name and the number circled on two 3x5" cards. One card was retained by an investigator and the second concealed in the participant's pocket. The PDD examination procedure was briefly explained to the participant. It was emphasized that the participant should not reveal which number he had circled during the PDD examination. It was further emphasized that the participant should make every attempt to remain relaxed, even if he felt himself begin to react (increased heart rate, perspiration on hands, tightening of occlusive cuff) during the examination. The participant was then escorted to the examination room and introduced to the examiner.

The examiner greeted each participant, then reviewed the biographical/medical questionnaire with the participant to ensure its accuracy. No other pre-test questions were asked by the examiner. The examiner then briefly explained the sensors, procedures, and theory of PDD. The examiner explained that the polygraph simply measured the participant's physiological reactions — and not deception per se. It was further explained that the participant's physiological responses were likely to change during deception. It was suggested that fear of detection during deception altered the normal physiological response pattern and that these changes may be evident in the recorded physiological data. The examiner described this response as being similar to the fight-or-flight reaction used to describe a fear response during military training.

The examiner reviewed the questions to be asked during data collection with the participant by playing the computer recorded questions. If there were no further questions, the participant was then seated in the examination chair and the sensors were attached. Respiration was monitored using convoluted (pneumo) tubes placed around the upper and lower chest. Skin resistance was measured using electrodes placed, with paste, on the most distal phalanges of right hand index and ring fingers. Cardiovascular activity was monitored using an occiusive cuff placed over the brachial artery of the left arm. The pneumo tube vents were closed and the DC offsets for the pneumo and skin resistance on the custom built amplifier were adjusted to zero. The sensitivity of these recording channels was then adjusted on the polygraph. Next, the occlusive cuff was inflated to 90 mmHg, massaged to remove wrinkles, then deflated to 48 mmHg. The pressure was then adjusted, as necessary, to achieve a 2 mmHg pen deflection, between diastole and systole, on the sphygmomanometer. The custom built amplifier DC offset was then adjusted to zero to keep the signal within the range of the analog-to-digital converter, and polygraph

McClellan, AL) and was certified as a PDD examiner by the Department of the Army. He had administered approximately 500 field examinations during the five years prior to the study and was an instructor at the DoD Polygraph Institute.

Apparatus

Data were collected using a Lafayette (Lafayette, IN) Factfinder (Model 76740 /76741) polygraph equipment with three Cardio/Aux/Pneumo/GSR modules (Model 76477-G), one GSR module (Model 76480-G), and one electronic stimulus marker module (Model 76351-GET). A circuit was added to the electronic stimulus marker module to allow control of the marker via signals from a computer RS-232 serial port. Lafayette sensors were used to measure skin resistance (Model 7664), respiration (Model 76513-1G & 76513-2B), and cardiovascular activity (Model 76530).

A stimulus presentation micro-computer (Model 248, Zenith Data Systems, Chicago, IL), was used to replay questions throughout testing. The questions used throughout PDD testing were digitized and recorded to computer hard disk using a Sound Blaster board (Model 16ASP, Creative Labs, Inc., Milpitas, CA). A parallel port interface (Speech Thing, Covox, Inc., Eugene, OR), connected to a Radio Shack (Fort Worth, TX) integrated stereo amplifier (Model SA-155) and two speakers (Model Minimus-77) was used to present the questions. This system ensured that each question was presented with the same inflection, and at the same volume, each time it was asked.

Subjects' verbal responses were recorded on cassette tape using a Tascam Model 134 4-channel recorder (TEAC, Montebello, CA) and a lavaliere microphone (Model 570S, Shure, Evanston, IL) positioned mid-chest and held in place by a cord placed over the examinee's shoulders. The recorder was located in an adjacent room. Excerpt recording was controlled via the software running on the stimulus presentation computer. The stimulus presentation computer serial port and an in-house built interface for the cassette recorder were used for this purpose.

A DT2821 data acquisition board (Data Translation, Inc., Marlboro, MA), installed in a standard IBM compatible 486 computer, in conjunction with Canadian Speech Research Environment software (CSRE 4.0, University of Western Ontario, Elborn College, London, Ontario, Canada), was used to acquire and digitize the analog voice signals from audio tape. A TTE 411AFS anti-aliasing filter (TTE, Inc., Los Angeles, CA) set to an upper frequency cutoff of 5000 Hz was installed between the tape recorder output and the DT282I input during conversion of the audio responses from analog to digital format. The voice spectrograms and pitch tracks were printed on 8.5" x 11" paper using a Hewlett-Packard XL300 color printer. Software was written in-house for data reduction and display.

PDD testing was conducted in a carpeted, 11'6 x 12' partially sound-attenuated room. Each examination was recorded on video tape using two ceiling and one wall mounted video cameras. The examination was also monitored through a two-way mirror by a collaborator located in an adjacent room.

Subjects were seated in a Lafayette adjustable-arm subject chair (Model 76871, Lafayette, IN) during testing. The chair was positioned beside and slightly in front of the examiner's desk. This position allowed the examiner to monitor the examinee's movements but not vice versa. The polygraph was mounted in a double pedestal examiner's desk (Lafayette Model 76183). The stimulus presentation computer and monitor were on a table next to the examiner's desk and out of the examinee's sight during testing. The speakers, through which the questions were played, were located six feet behind, and one foot above, the back of the examinee's chair. The examinee's field of view, throughout testing, was limited to a wall of uniform color, a stationary video camera, and, above the video camera, a piece of paper with the numbers 60 through 66 and the word "NO" written on it.

Procedure

Participants were randomly assigned to the treatment or control groups, with the constraint that at least one volunteer from the time when unprocessed speech was presented to them. Using speech synthesis techniques, they found that identification accuracy dropped to 25% when pitch information within the raw speech waveform was smoothed. Their conclusion was that pitch perturbations in human speech were important to the transmission of emotional information, and that this was an "acoustic correlate of some phonetic or emotional event."

In another study focusing on pitch changes, Streeter, Krauss, Geller, Olson, and Apple (1977) found that subjects' average response fundamental frequency (FO) was higher when they were being deceptive than when telling the truth. In addition, they found that the magnitude of this difference was marginally greater when the deceptive act was stressful or arousing. Tolkmitt and Scherer (1986) reported that mean FO is less sensitive to stress than FO floor, and that FO floor may be a better indicator of stress (FO floor rises when arousal increases). FO floor was defined as the final FO value of a speaker's declarative statement.

Another method, commonly referred to as PSE (psychological stress evaluation), has met with varying degrees of success (Barland, 1978; Brenner, Branscomb, & Schwartz, 1979), but has never been widely accepted by PDD examiners as a reliable tool. This lack of acceptance may largely be due to the fact that PSE was meant to replace rather than augment the standard polygraph, and by itself may not provide sufficient information for confident judgment. A major drawback is that PSE appears to rely solely on changes in the FM (frequency modulation) component of speech, most often referred to as microtremor, for the detection of deception. The reliability of the relationship between voice microtremor and autonomic reactivity has not been well established. Evidence from controlled studies shows that voice stress analyzers fail to yield deception detection rates above chance levels (Horvath, 1982),

The present study was designed to examine the verbal responses of subjects to determine if features within the acoustic components are related to deception. Analyses were performed on the pitch contours (time domain) and spectral energy patterns (frequency domain) of subjects' voice responses during a peak-of-tension (POT) numbers test. The FM component, mean dominant (fundamental) pitch frequency, response duration, and mean response intensity of deceptive and truthful "no" responses were examined. Changes in the magnitude and rate of the FM component were also expected. In the frequency domain it was expected that deceptive responses would result in a spectral energy pattern shift when compared to non-deceptive responses.

Method

Data Collection

The data used in this study were collected during a repeated measures study (Dollins, Cestaro, & Pettit, 1994). A complete description of the procedures used throughout data collection is included for accuracy, though many of the procedures were not directly related to this voice analysis study.

Subjects

Forty-four, native English speaking, healthy males [mean age (SD) = 29.2 (7.8) years, range = 19 to 47] participated in this study. Volunteers were civilian or military Department of the Army employees and were not paid for their participation. Thirty-nine of the volunteers had never participated in a PDD examination before. The remaining five volunteers had not participated in a PDD examination within the last two years. Thirty-five of the volunteers reported themselves to be medication free. The remainder were ingesting pain/relaxant (3), anti-inflammatory (l), antibiotic (2), or antihistamine (3) medication. Females did not participate in the repeated measures study because of possible variations in skin resistance (over time) caused by hormonal secretions associated with the menstrual cycle. The data of 16 subjects were excluded because response amplitude was too low, leaving 28 subjects' data for analysis. Six of these subjects were using one of the above-mentioned medications.

Examiner

All PDD examinations were conducted by the same examiner. The examiner had completed training at the Department of Defense (DoD) Polygraph Institute (Fort

An Analysis of Voice Responses for the Detection of Deception

Victor L. Cestaro and Andrew B. Dollins

Abstract

This study was designed to examine the feasibility of using audio pitch analysis and spectrum decomposition techniques to aid in the detection of deception following a numbers test. Audio recordings were made of 44 male subjects' responses during a peak-of-tension (POT) test. A Lafayette field polygraph was used to collect respiration, cardiovascular, and electrodermal responses for manual evaluation. Half of the examinees were programmed "deceptive" and half were programmed "truthful". The audio recordings of the subjects' responses were analyzed off-line using pitch and spectral analysis software to examine differences between truthful and deceptive "no" responses. Useable voice recordings were obtained from 28 of the original 44 subjects. No significant differences were found between the two groups on individual measures of pitch variation, response duration, or mean response energy. There was a significant concurrence rate ($p < .01$) between decisions made by pitch/energy analysis and an examiner based on analysis of the test data. Significant differences were found between the number of correct decisions made by the examiner (79%) and by pitch/energy analysis (37%). However, no significant differences were found between the number of false positive decisions made by the examiner and by pitch/energy analysis (35% versus 29%).

Standard psychophysiological detection of deception (PDD) tests and procedures have historically used measures of autonomic nervous system reactivity to differentiate between deceptive and non-deceptive subjects. Changes in skin resistance, breathing rate, and cardiovascular activity in response to questions requiring a "yes" or "no" answer have been the most common measures. In most cases, decisions are based on analysis of the physiological data recorded using four polygraph channels (cardiovascular, electrodermal, and two respiratory channels). There have been no additional channels added to the traditional polygraph since its inception as a tool designed for the PDD. However, various attempts have been made in the past to detect deception using voice stress analysis (O'Toole, 1975). Interest in this method was reported more than five decades ago in a study conducted by Fay and Middleton (1941) who relied on human judgments of voice responses to determine truth or deception. Forty-seven subjects were told to answer a series of ten questions either truthfully or untruthfully. Instructions to lie or tell the truth were presented immediately before each response, and subjects' responses were judged by a panel of 60 observers. Correct judgments were at or near chance levels, with judgments of "lie" answers slightly better than truthful answers (60.99% vs. 50.05%).

Using more sophisticated techniques, Motley (1974) examined extracted pitch information from voice responses in an attempt to detect involuntary (autonomic) manifestations of stress related to deception. Twenty female subjects were instructed to respond "no" to a series of questions related to a number picked prior to the experimental session. Analysis of recorded responses examined intensity, fundamental frequency, duration, formant structure, and harmonics. The only difference found between truthful and deceptive responses was in the response duration measure ($p < .01$). A second procedure in this experiment showed that acoustic cues associated with deception were not detectable by the unaided ear at better than chance levels, which lends support to the results obtained by Fay and Middleton (1941).

Other investigators have demonstrated an interest in the pitch component as an indicator of emotional content in speech (Lieberman, 1961; Lieberman & Michaels, 1962). Lieberman and Michaels (1962) stated that observers were able to correctly identify specific emotional states of subjects 85% of

Cestaro, V.L. (1996a). A comparison between decision accuracy rates obtained using the polygraph instrument and the computer voice stress analyzer (CVSA) in the absence of jeopardy. *Polygraph, 25,* 117-127.

Cestaro, V.L. (1996b). A test of the computer voice stress analyzer (CVSA) theory of operation. *Polygraph, 25,* 101-116.

Florida Department of Law Enforcement. (1993, October). *Review of literature regarding voice stress analysis.* A report compiled by the Office of Inspector General FDLE, Tallahassee, FL.

Horvath, F. (1982). Detecting deception: The problem and reality of voice stress analysis. *Journal of Forensic Sciences, 27,* 340-351.

National Institute for Truth Verification. (1990). *Journal of Continuing Education (Vol. 10, No. 6).* West Palm Beach, FL: Author.

National Institute for Truth Verification.(1994). *Journal of Continuing Education (Vol. 12, No. 1).* West Palm Beach, FL: Author.

National Institute for Truth Verification. (1995). *Journal of Continuing Education (Vol. 12, No. 3).* West Palm Beach, FL: Author.

U.S. Congress, Office of Technology Assessment. (1983). *Scientific validity of polygraph testing* (OTA-TM-H-15). Washington, D.C.: U.S. Government Printing Office.

Appendix A

CVSA Relevant Test Questions (MZOC)

IR	1.	
C	2.	
IR	3.	
R	4.	Do you know who took that $100 bill from that metal box?
IR	5.	
R	6.	Did you take that $100 bill from that metal box?
IR	7.	
C	8.	
IR	9.	
R	10.	Do you suspect anyone of taking that $100 bill from the metal box?
IR	11.	
R	12.	Do you know who took that $100 from that metal box?
IR	13	
R	14.	Did you take that $100 bill from that metal box?
IR	15.	

indicated that with N =100 (50 per group collapsed across programming [guilty, innocent] and an expected effect size of 0.20, power = .99 (p = .05). This means that, under the test conditions used in this study, there is a .99 probability that an effect of .20 greater than chance would have been detected, had one existed.

While every attempt was made to emulate the subject programming procedures reported in other studies, it is possible that the procedures used did not elicit physiologic responses during deception. While, in our opinion, it is unlikely that the low accuracy rates obtained are due to problems with the mock crime scenario, it is a possibility.

The test procedures incorporated in the study were the same as those used in field examinations, and all seven examiners (administering and evaluating) were trained and certified by the equipment manufacturer. All examiners had practical field experience in the pre-test, in-test, test analysis, and post-test phases of CVSA examination administration, and the administering examiners were permitted to conduct the examinations as learned in certification training conducted by the NITV. Examinations were monitored by a CVSA instructor from the NITV. The statistically significant decision concurrence rate, as shown by the results of the interrater reliability tests, seems to provide some level of confidence that the scoring methods employed among examiners were consistent. However, from a practical viewpoint, examiners obtained majority decision agreement on less than half of the subjects, and unanimous agreement on about one quarter of the subjects tested. The lack of instrument sensitivity to the measure(s) of interest impacted on the ability of examiners and evaluators to accurately and consistently discriminate between truthful and deceptive responses when assessing the subject's test charts.

In summary, although there is evidence to support the basic electrical theory of operation of the CVSA (Cestaro, 1996b), the instrument failed to function in a manner that would allow examiners to discriminate between truthful and deceptive responses from test subjects. Further research should examine the effects of increased levels of stress on subjects' responses to determine if there is a correlation between stress levels and instrument display characteristics. Although the CVSA instrument is purported to detect stress in human speech, there is still no unambiguous evidence to support that claim.

Acknowledgments

The authors wish to express sincere gratitude to Detective Gary Noel and Robert Neubauer, and Investigators Terry Kilgore and Bernie Erwin for serving as the examiners in this study. Special appreciation goes to Special Agent Bob Tippett, Florida Department of Law Enforcement, for his technical advice on the Computer Voice Stress Analyzer (CVSA), and for thoroughly reviewing the study protocol. Also, special thanks to Brenda Smith, Charlene Stephens, Kathy Harris, Linda Landy, Johan Harrison-Woodard, and Sarah Tidwell.

This project was funded by the Department of Defense Polygraph Institute as DoDPI95-P-0016. The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. For reprints, contact Dr. Victor Cestaro, DoDPI, Bldg. 3195 and 13th Avenue., Ft. McClellan, AL 36205.

References

Ansley, N. & Garwood; M. (1984) . The accuracy and utility of polygraph testing. *Polygraph, 13,* 3-131.

Certified Examiners Course Manual (1995) (Available from the National Institute for Truth Verification, West Palm Beach, FL.)

the subject and the CVSA instrument was calibrated for the subject's voice level. The examination proceeded using the accepted CVSA format for the Modified Zone of Comparison test and appropriate test questions.

The CVSA examiners conducted three examinations. The chart from the first examination was not evaluated, in accordance with NITV scoring procedures (*Certified Examiners Course Manual*, 1995). The second and third charts were numerically scored, and categorized as Deception Indicated (DI) or No Deception Indicated (NDI). In addition, all examinations were recorded on video/audio VHS tape for off-line analysis. When the examination was completed, the subject was escorted back to the briefing room for subject debriefing.

Scoring

Before data reduction and analysis, the original examiners independently evaluated each graphic recording. Based on their scoring they were asked to make a diagnosis of either DI or NDI. CVSA procedures do not allow for inconclusive determinations. The examiners' scores and decisions were not written on the charts. The decision for each subject was written by each examiner on a scoring sheet maintained by the examiner. All charts were marked only with the date of the examination and subject number. Charts blind scored by the three evaluators had all subject numbers removed, and were randomly coded.

Data Analysis

The dependent measure for accuracy was the number of correct decisions made regarding deception and non-deception. Interrater agreement was determined by comparing the decision made for each subject among the evaluators, irrespective of the accuracy of the decision. Analyses included a 2 x 2 chi-square analysis of programming *vs.* decision, and a test of the significance of proportions of DI and NDI decisions when compared to chance expectancy (0.50). An in-house program using common signal detection theory (SDT) procedures was used to assess instrument sensitivity. Scoring reliability (in the form of interrater agreement) was assessed by a multiple rater kappa statistic (Fleiss, 1981).

Results

Evaluators made correct decisions on 163 of 327 charts (109 subjects x 3 evaluators), obtaining an overall accuracy of 49.8% (z = -.05, p = .96), with a range of 45.9% to 54.1%. Their accuracy ranged from 54.5% to 63.6% for DI decisions, and 35.2% to 53.7% for NDI decisions. Administering examiners did slightly worse, achieving an overall accuracy of 48.6% (z = .21, p = .84), with an accuracy range of 33.3% to 55.6%. Their DI decision accuracy ranged from 38.5% to 66.7%, and their NDI decision accuracy range was 13.3% to 66.7%. No examiner obtained a combined (DI and NDI) accuracy rate significantly different from chance, nor were the results of chi-square analyses significant. Application of SDT to the data showed that overall instrument sensitivity was low. The noise and signal + noise distributions were completely overlapped, with the criterion line (beta) positioned near the means of the overlapped distributions, indicating nearly equal probability for DI or NDI decisions (d'=0, beta = 1.01). Interrater reliability (mean proportion of agreement) for decisions rendered was conducted by three blind score evaluators. These evaluators obtained a correct unanimous agreement rate of 26%, and a correct majority (2 of 3) agreement rate of 46%. Interrater reliability for all decisions rendered by the evaluators was high (kappa = .33, SE = .055, p < .0001). These evaluators obtained a correct unanimous agreement rate of 26%, and a correct majority (2 or 3) agreement rate of 46%.

Discussion

As shown in a previous study (Cestaro, 1996a), the sensitivity of the CVSA is low when used in a low or no-stress situation, such as that encountered during a typical laboratory study. The CVSA manufacturer claims that stress related to deception can be detected reliably by the instrument, and that stressful and non-stressful responses can be differentiated by trained operators. However, in this study, evaluators and administering examiners were not able to distinguish between deception and non-deception at rates better than chance levels of accuracy (50%). Prior to conducting the study, a power analysis of the proportion test for accuracy

In 1993, the Inspector General of the Florida Department of Law Enforcement (FDLE, Tallahassee, FL) released a position paper recommending that FDLE prohibit the use of voice stress analysis as an investigative tool because of the lack of scientific evidence supporting its validity. Since the CVSA records physiological data from a response system (the voice) that the current polygraph is incapable of recording, it is possible that the combination of instruments and processes (polygraph and the CVSA) could increase the accuracy and reliability of the detection of deception. The purpose of this study was to evaluate the CVSA and its associated procedures to determine its efficacy in detecting deception.

Method

Subjects

One hundred nine subjects were recruited from a local contract agency and randomly assigned to deceptive and non-deceptive groups. Volunteers were male or female, literate, between the ages of 19 and 65 years, and had a minimum of a high school diploma or GED. Each subject was determined to be in good health and slept at least six hours the evening before testing.

Apparatus

Four Computer Voice Stress Analyzers (National Institute for Truth Verification, W. Palm Beach, FL) were used to record and display voice response data on paper charts. Lapel microphones (Radio Shack, Fort Worth TX, Model 33-3003) were used for supplying subjects' verbal responses to the input jacks of the CVSAs.

Examiners

Four CVSA examiners, trained and certified by NITV, conducted the examinations. The examiners were blind to subject programming. The CVSA tests were also independently blind scored by three trained and certified CVSA examiners, hereinafter referred to as evaluators.

Procedures

Upon arrival at the Department of Defense Polygraph Institute (DoDPI) testing site, each participant was escorted by a research team assistant to the DoDPI library and asked to read a brief description of the research project. Subjects were programmed in groups of four, two groups in the morning and afternoon. Individuals willing to participate in the study were asked to read and sign a volunteer agreement affidavit. A brief biographical/medical questionnaire was completed to ensure that each participant was in good health and not taking medication that could interfere with examination results. Research team assistants than began programming deceptive and non-deceptive subjects according to the scenario instructions. All subjects were then given their appropriate written instructions. Random assignments of subjects to groups were made before the actual experiment. Nearly half of the subjects (n = 55) were assigned to the deceptive group and participated in taking $100 from a metal box located in a scenario room. The remaining subjects (n = 54) did not participate in the scenarios nor did they have knowledge of the mock theft.

Deceptive subjects were instructed to proceed to the scenario room and to remove the $100 bill from an open metal box located on a table in the scenario room. Each deceptive subject was told to hide the $100 bill on their person. Additionally, they were instructed to lie to the examiner about taking $100 from the metal box and having the money on their person. Next to the metal box was a 3" x 5" card with each deceptive subject's examiner room assignment. Non-deceptive subjects also entered the scenario room and picked up a 3" x 5" card with their examiner room assignment. However, the metal box containing the $100 bill as removed before non-deceptive subjects entered the scenario room. They had no knowledge of the theft, and were instructed to answer questions truthfully during the examination.

CVSA examiners conducted the pre-test interview as described in the NITV *Certified Examiners Course Manual* (1995). The relevant questions used were the same for all subjects (Appendix A). The control and irrelevant questions used were developed by each examiner, based on the rules of test question formulation taught in the NITV Certified Examiners Course. All test questions were reviewed with the subject before testing began. The lapel microphone was placed on

Effectiveness of Detection of Deception Examinations Using the Computer Voice Stress Analyzer

Michael J. Janniro and Victor L . Cestaro

Abstract

The accuracy of the Computer Voice Stress Analyzer (CVSA) instrument and associated processes for the detection of deception was assessed using a mock theft scenario. One hundred nine subjects were randomly assigned to two groups and given detection of deception examinations using a CVSA instrument. Subjects in one group were programmed deceptive and participated in taking $100 from a metal box located in a scenario room. The non-deceptive group did not participate in the scenarios nor did they have knowledge of the mock theft. Four trained and certified CVSA examiners conducted the examinations using a CVSA technique called the Modified Zone of Comparison test. CVSA test chart evaluators, who had not taken part in the study and who were blind to subject programming, obtained an overall accuracy of 49.8% (z = -.05, p = .96) . Administering CVSA examiners correctly identified 53 of the 109 (48.6%) subjects as either deceptive or non-deceptive (z = -.21, p = .84). More deceptive subjects were correctly identified by examiners than non-deceptive subjects (32 of 55 vs. 21 of 54). However, decisions were not significantly different from chance in either case.

Key words

Computer Voice Stress Analyzer, CVSA, detection of deception, voice stress

In 1971 Dektor Counterintelligence and Security, Inc., (Savannah, Georgia), developed a device for detecting stress, which they called the Psychological Stress Evaluator (PSE). The National Institute for Truth Verification (NITV) *Certified Examiners Course Manual* (1995) states that the PSE detects subaudible micro tremors in the human voice, and that analysis of these stress related tremors has great utility for the detection of deception. Soon afterwards, advertisements in popular magazines, newspapers, and trade journals began comparing the accuracy and utility of the polygraph to voice stress analyzers (NITV, 1990; NITV, 1994; NITV, 1995). Claims have been made in newspaper articles that the CVSA is easier to use and more accurate than the polygraph (NITV, 1990, p. 18).

The PSE has recently been supplanted by an instrument called the Computer Voice Stress Analyzer (CVSA) manufactured by the NITV. Although the theoretical physiological basis of monitoring subaudible micro tremors is unchanged from the PSE, instrument design changes and ease of use are making the CVSA a popular tool. Periodic publications of the NITV's *Journal of Continuing Education* (e.g., 1990) include several newspaper articles pronouncing the CVSA's effectiveness and acceptance by many police departments. Most testimonials cited in NITV's journals, regarding the efficacy of the CVSA, stress its utility in obtaining admissions and confessions. However, the manufacturer does not provide evidence of controlled laboratory studies that would support the high accuracy rates (97-100%) routinely claimed (NITV training registration form). Furthermore, no explanations are provided for how these accuracy rates are determined.

The effectiveness of the polygraph, has been the subject of a number of controlled scientific studies over the years (Ansley & Garwood, 1984; U.S. Congress, Office of Technology Assessment, 1983). According to Horvath (1982), many well-controlled laboratory studies and field studies support the polygraph and its associated procedures and processes. Horvath argues that even the most severe critics agree that the findings show an accuracy that justifies the use of polygraph testing under certain conditions. However, the CVSA and its testing procedures and processes have not been subjected to the same vigorous scientific evaluation as the polygraph and its procedures.

About the Authors

Donald J. Krapohl is a researcher with the Department of Defense Polygraph Institute, and a regular contributor to this journal. He can be contacted at krapohld@jackson-dpi.army.mil.

Dr. Andrew H. Ryan is the Chief of Research, US Department of Defense Polygraph Institute. He can be reached at ryana@jackson-dpi.army.mil.

Kendall W. Shull recently retired as Chief of the FBI's Polygraph Unit. He is now in private practice as Kendall Investigations and Polygraph Services in Knoxville, Tennessee. He can be contacted at KendallShull@aol.com, or (865) 742-7744.

This article is being published concurrently in the *Policy Review* newsletter, Vol. 13, No.4, (2002).

The views expressed in this article are those of the authors, and do not necessarily represent those of the Department of Defense or the US Government.

that voice could become one channel in the next generation of lie-detection instrument that might also include brain waves, eye movement, thermal imaging, remote sensing, or some technology that does not yet exist. Though none of the current voice analysis technologies are valid for detecting deception, the US Government's continuing investigation in this area might one day find one that works, with the goal of better protection of our communities and our nation.

References

Brenner, M., Branscomb, H., & Schwartz, G.E. (1979). Psychological stress evaluator: Two tests of a vocal measure. *Psychophysiology, 16*(4), 351-357.

Brown, T.E., Senter, S.M, & Ryan, A.H. (In press). Ability of the Vericator™ to detect smugglers at a mock security checkpoint. Abstract.

Cestaro, V.L. (1996). A comparison between decision accuracy rates obtained using the polygraph instrument and the Computer Voice Stress Analyzer (CVSA) in the absence of jeopardy. *Polygraph, 25*(2), 117-127.

Cestaro, V.L., & Dollins, A.B. (1996). An analysis of voice responses for the detection of deception. *Polygraph, 25*(1), 15-342.

DoDPI Research Division Staff, Meyerhoff, J.L., Saviolakis, G.A, Koenig, M.L., Yourick, D.L. (2000). Physiological and biochemical measures of stress compared to voice stress analysis using the Computer Voice Stress Analyzer (CVSA). Report No. DoDPI98-R-0004. Fort Jackson, SC.

Hollien, H., Geison, L., Hicks, J.W. (1987). Voice stress analysis and lie detection. *Journal of Forensic Sciences, 32*(2), 405-418.

Horvath, F.S. (1978). An experimental comparison of the psychological stress evaluator and the galvanic skin response in detection of deception. *Journal of Applied Psychology, 63*(3), 338-344.

Horvath, F.S. (1979). Effect of different motivational instructions on detection of deception with the psychological stress evaluator and the galvanic skin response. *Journal of Applied Psychology, 64*(3), 323-330.

Janniro, M.J., & Cestaro, V.L. (1996). Effectiveness of detection of deception examinations using the Computer Voice Stress Analyzer. *Polygraph 27*(1), 28-34.

Lynch, B.E., & Henry, D.R. (1979). A validity study of the psychological stress evaluator. *Canadian Journal of Behavioural Science, 11*(1), 89-94.

Palmatier, J.J. (1999). The Computerized Voice Stress Analyzer: Modern technological innovation or 'the emperor's new clothes'? *GP Solo & Small Firm Lawyer, 16*(4), 42-45.

Sommers, M.S., Brown, T.E., & Ryan, A.H. (In press). Evaluating the reliability and validity of Vericator™ as a voice-based measure of deception. Abstract.

Timm, H.W. (1983). The efficacy of the psychological stress evaluator in detecting deception. *Journal of Police Science and Administration, 11*(1), 62-68.

Vericator User Manual (2000). Integritek Systems, Inc.: Tampa, Florida.

Waln, R.F. Downey, R.G. (1987). Voice stress analysis: Use of telephone recordings. *Journal of Business and Psychology, 1*(4), 379-389.

It had been shown that the CVSA did not perform well in low-jeopardy scenarios, and it therefore became important to test it in settings in which the outcome was more meaningful to the examinee. In a mock crime study Janniro and Cestaro (1996) again evaluated the accuracy of the CVSA to detect deception. One hundred nine subjects were tested; half were asked to commit a realistic and engaging mock crime while half did not participate nor had knowledge of the mock theft. CVSA examiners conducted and scored the exams in accordance with NITV procedures. Charts were blind-scored by three other CVSA evaluators. The variable of interest was the number of correct decisions, with chance accuracy set at 50%. Blind CVSA evaluators made correct decisions on 49.8% of the cases, while the testing CVSA examiners achieved 48.6% accuracy. These accuracy rates were not different than chance. The authors concluded in this laboratory paradigm that, though the examiners consistently employed the NITV scoring methods, the CVSA sensitivity to detect lies was low.

The last DoDPI study with this device was a collaborative project conducted jointly by DoDPI and the US Army Walter Reed Hospital (2000). The project examined the capabilities of the CVSA in a well-understood and controlled stressful interview model (US Army Soldier of the Month Board). In this study, voice responses before, during, and after the interview were transferred to CVSA charts for blind scoring by CVSA evaluators. In addition, other indices were recorded before, during, and after the interview using validated medical measures of physiological stress. These included heart rate, arterial blood pressure, and plasma ACTH. Salivary cortisol measures were made before and after the interview. The results showed that the interview paradigm elicited stress at significant levels, as indicated by the medical markers of stress. Results for the CVSA did not correlate with the medically confirmed stress at any level, neither low nor high. In addition, inter-CVSA examiner agreement proved to be low. The authors conclude that the CVSA analysis of voice features does not reflect well-validated tonic responses to acute stress. In other words, whatever the CVSA may record, it is not stress. The makers of the CVSA have since suspended cooperation with federal research with their product, and have required some new buyers of their equipment to agree not to participate in government-sponsored validity research.

The studies outlined above cast strong doubt on the ability of micro-tremor analyses to detect deception better than chance. A new product, Vericator by Integritek Systems, Inc. of Tampa, Florida, has recently been introduced that claims to extract information from the entire vocal signal to produce decisions (2000). It has been marketed in a manner that emphasizes its flexibility and utility across a wide range of situations and circumstances. DoDPI's interest has grown in the possible use of the Vericator as a tool that could facilitate the work of inspectors at security checkpoints (e.g., US Customs), a setting where polygraph examinations are not practical. DoDPI commenced a two-site comprehensive study to assess the accuracy of the Vericator to detect deception. Detection rates at both sites, which involved very realistic and stress-inducing laboratory paradigms, were quite disappointing and did not exceed chance (Brown, Senter, & Ryan, 2002; Sommers, Brown, & Ryan, 2002.)

Conclusion

Over the last 30 years other researchers outside of the government have also researched voice stress for lie detection, and published their findings in scientific journals. The general conclusion has been that the accuracy is modest to poor for a handful of experimental approaches, and uniformly poor for those relying on the micro-tremor (see www.voicestress.org for a summary of the available research). This does not prevent some as-yet untried analytical approach from someday yielding a valid voice lie-detector, and the Government is still aggressively seeking such a capability for the important advantages it would afford. As a practical consideration, the poor validity for the current voice stress technology should provide a caveat to agencies considering adding voice stress to their investigative toolboxes.

The controversy over the use of voice stress analysis will surely continue for years to come. Additional research offers the prospect

If such a relationship between these micro-tremors and deception were empirically sound, government security professionals and law enforcement would have a powerful new tool, not only to replace the polygraph, but for applications where the polygraph cannot be used. A review of the current research is presented here to bring the reader up to date with the findings.

The first significant commercially available product to analyze vocal signals was the Psychological Stress Evaluator (PSE), introduced in the early 1970s by Dektor Counterintelligence and Security, Inc. of Springfield, Virginia. All analyses were conducted off line, using an audio recording of the examinee's voice taken during a structured test. Quick and inexpensive, hundreds of PSEs were sold, though they never achieved the acceptance enjoyed by the polygraph, due in large part to the lack of supporting evidence that it could actually detect deception (Brenner, Branscomb, & Schwartz, 1979; Hollien, Geison, & Hicks, 1987; Horvath, 1978, 1979; Lynch & Henry, 1979; Timm, 1983; Waln & Downey, 1987).

In the late 1980s, the National Institute for Truth Verification, Inc. (NITV) of West Palm Beach, Florida produced what they termed a computer voice stress analyzer and trademarked it as CVSA. The CVSA is marketed as a convenient replacement for the polygraph. Like the PSE, the CVSA analyzes micro-tremors in the vocal signal, but unlike the PSE, the CVSA provides real-time graphical outputs or charts that examiners can score. Sales of the CVSA have been brisk in recent years, easily overshadowing other brands of voice stress devices. The US Government purchased a small number of units, and trained a few personnel, but after field trials with the devices did not meet expectations, the equipment was discarded. Widespread use of the CVSA in the law enforcement sector, combined with the Government's continuing interest in new lie detection methods, prompted DoDPI to conduct and sponsor a number of studies to answer two important questions. First, can micro-tremors in the vocal signal be used effectively to detect deception? Second, how does this compare to the current gold standard, the polygraph?

Cestaro and Dollins (1996) examined the utility of using the vocal responses of subjects during a low-stress test involving the examinee concealing a number. Parameters examined were spectral energy distribution of the voice response, fundamental frequency, response energy, response duration, and pitch variations around the fundamental frequency. No significant relationships were found in the voice data for vocal components and deceptive answers. The authors concluded that none of these parameters, in isolation, was a reliable and valid discriminator of truth and deception. However, they left open the possibility that multiple measures extracted from pitch information might be useful as indicators of deception.

A two-part study by Cestaro (1996) was conducted because controlled lab research to test the validity or reliability of the CVSA instrument or the techniques employed in its use had not been conducted. The first part was designed to determine whether the CVSA detects micro-tremors in the fundamental frequency of presented signals as the manufacturer claimed. The second part was designed to determine whether accuracy rates obtained using the CVSA differed from those using the traditional polygraph. The results of part one indicated that the CVSA functioned electrically according to the manufacturer's theory of operation. Changes in the frequency of the input signal caused deflections of the CVSA display in proportion to the frequency of the input signal. Because the study demonstrated that the CVSA recorded what it purported to be recording, the issue of decision accuracy was then ready to be investigated.

In the second part, the CVSA was compared to the polygraph, again in a low-stress lab study. Forty-two subjects were tested with both instruments. The difference in accuracy between the polygraph and CVSA was significant: polygraph decisions were significantly greater than chance, while the CVSA decisions were not. The authors concluded that poor instrument or procedure sensitivity of the CVSA was the cause for the lack of accuracy.

Accuracy may not be all that important in some applications, of course, such as when the device is used only as an adjunct to an interrogation. In an interrogation setting, a sophisticated-appearing machine in the room that is represented as a lie-detector may offer the interrogator a psychological wedge to encourage more candor from the suspect. Whether the machine really detects anything is secondary, so long as the suspect believes it works. Repeated judicial decisions have supported the use of ploys and trickery by law enforcement to help obtain a confession, so long as the tactic is not so coercive as to "shock the conscience." Given the non-intrusiveness of current voice stress products, it seems likely that they would likely withstand that test.

The Costs

Low validity is not without drawbacks, some potentially severe. Poor accuracy could have profound consequences for a department if investigative decisions are based on the outcome of the voice stress examination. Precious manpower resources could be misdirected, or a criminal could escape while another citizen is wrongly pursued, affecting not only public safety, but community confidence, as well. Also, the use of the devices in a surreptitious mode raises very imposing legal questions, issues that are beyond the scope of this article, but are especially important when placed in the context of the devices' unimpressive accuracy.

Of more immediate concern are those instances where departments use voice stress tests in the hiring process for new officers. In 1999, the American Bar Association published an article (Palmatier, 1999) which indicates that the use of a voice stress device for hiring decisions may constitute a violation of the Equal Employment Opportunity Commission rules, and that the departments could find themselves in legal peril if they use them in this manner. It is because of the twin problems of validity and potential litigation that voice stress technology has played a limited roll in state and local law enforcement, and none at the federal level.

Federal Research

Recognizing its possible applications, the US Government's interest in voice stress technology can be traced back at least to the 1960s. A number of government agencies independently investigated the potential of this method of lie detection, but since the mid-1980s the task has fallen largely to the research facilities at the US Department of Defense Polygraph Institute (DoDPI). In its charter, DoDPI is responsible for providing research in new concepts and technologies with relevance to the detection of deception. Though DoDPI's research into alternative methods, including voice stress, has been ongoing for over 10 years, those efforts have taken on new emphasis since the terrorist attacks on New York City and Washington D.C. Voice stress is currently one of the hot topics, and DoDPI has conducted or collaborated in several studies on voice stress devices that can provide answers to agencies and departments weighing the potential costs and benefits of fielding them.

The premise of all voice lie detectors is that certain pitch parameters, allegedly associated with certain nervous system activities, are not under voluntary control. Voice stress device marketers suggest that there is an inaudible component in the vocal spectrum, called the micro-tremor, which changes during stress. Micro-tremors are oscillations in the FM component of the voice, in the range of 8 to 14 cycles per second, and purportedly are markers for the stress associated with the act of deception. According to descriptions by the manufacturers, micro-tremors are normally seen in relaxed, natural speech. Their disappearance signals stress, with the inference that the speaker is uncomfortable with what he is saying. In the field, a voice stress technician asks a structured series of questions during the voice stress examination, some questions relating to the crime and others being neutral. By noting the presence or absence of micro-tremors on the crime questions, a decision is rendered regarding the examinee's truthfulness.

Voice Stress Devices and the Detection of Lies

Donald J. Krapohl, Andrew H. Ryan, and Kendall W. Shull

Introduction

Throughout history, one of the most common and difficult problems for law enforcement officers has been determining whether a potential suspect is lying or telling the truth. Many an investigation has stalled, been diverted, or failed, because the statements of a suspect couldn't be verified, regardless of how much effort the investigators expended. Law enforcement has often turned to science with this problem, and science has tried to provide a solution. In the last 100 years several techniques and devices have been proffered to police to help them separate fact and fiction in suspects' stories, from interrogation drugs to word association tests, reaction time tests to interpretations of body postures, eye contact and gestures, - even a machine that registers hand trembles. The polygraph notwithstanding, no lie-detection method has enjoyed much longevity. In the last 25 years, however, newer methods have been introduced that are purported to offer more convenience, accuracy, and utility than even the polygraph. These are the voice stress devices, and by now most police agencies have seen or heard about them. They come with modest up front costs, making them very attractive to cash-strapped departments. Marketers of voice stress instrumentation portray their product as an important tool to hundreds of local police agencies, that they solve crimes quickly, and help in the selection of qualified police candidates.

The prospect of detecting lies in a speaker's voice has intuitive appeal. A common experience for most people is to have spotted a fib simply by noting a change in the tone of a speaker's voice. It would seem logical then, that with the help of computers and advanced technology, the lies of suspects could similarly be detected. If a device could find that special something that happens in the voice when a person is trying to deceive us, it could be a boon for the criminal investigation process, as well as for a variety of other uses such as business negotiation, confirmation of treaty compliance, airport security, and insurance claim verification, to name a few. The purpose of this article is to review what is known about voice stress, and to assess to what degree this technology can provide a reliable means for detecting deception.

The Benefits

Voice stress devices offer several potential advantages over the standard polygraph, the reigning lie-detection technology. The training time to operate a voice stress device is less than that for polygraph training, and there are no academic prerequisites to receive that training. Very low training and education requirements can save taxpayer money, and put the devices in the hands of more officers. The voice stress examinations themselves take little time, averaging about 45 minutes per session, or about a half or third of the time needed for a typical polygraph examination. There are no sensors placed on the body, only a small microphone clipped to the examinee's clothing. Because only the voice is used, the examinee need not even be present during the examination. A recording from a remote location or time can be processed with the equipment. Not only might this be more convenient than transporting the suspect to the voice stress technician, but it also opens the door to surreptitious processing of previously recorded voices.

Though thousands of the devices have been sold over the years, a far smaller number remain in service after a few years. Despite convenience and low cost, there are problems with voice stress devices that the product manufacturers have not yet overcome. The most pressing shortcoming appears to be the level of accuracy these machines deliver. As will be taken up later in this article, the track record of voice stress analysis in careful empirical studies has been lackluster. This has forced promoters to rely heavily on personal testimonials as evidence of accuracy.