Executive Summary
Voice Stress Analyzer Instrumentation Evaluation

H. Hollien and J.D. Harnsberger

This project had two primary goals: 1) Generate a database of speech materials to permit the testing of any present or future voice stress analyzer (VSA) and 2) Apply these speech materials in the evaluation of two VSA devices, the National Institute for Truth Verification's (NITV) Computer Voice Stress Analyzer (CVSA) and Nemesysco's (distributed in the United States by V, LLC) Layered Voice Analysis (LVA). These speech materials were generated by two types of experiments, *laboratory-based* and *field-based*. The laboratory-based studies provided a range of scenarios that varied stress, deception with jeopardy and speaker intention; they included:

1. Neutral, unstressed utterances (baseline material of several types)
2. Deceptive speech produced under low jeopardy
3. Deceptive speech produced under high jeopardy (high stress)
4. Truthful utterances spoken under high stress conditions
5. Simulated high-stress speech (the subject is actually experiencing low stress)

The laboratory materials involving jeopardy/stress were elicited by two methods. First, all subjects were drawn from a spectrum of the population involving age (18-55 years) and socioeconomic background; they also were selected from groups that were known to hold very strong personal views about some issue (e.g., religion, politics). They then were recorded uttering statements that contradicted their strongly held personal views on the selected issue, while under the impression that their friends and/or other peer groups would hear their utterances. In addition, subjects were instructed to produce these lies in a speaking style that strongly suggested that they believed them. Second, subjects were conditioned to respond to the highest level of electric shock that he/she could tolerate. They were told that they would receive a shock whenever they produced a neutral sentence appearing in the middle of the passage. In both methods, the induction of stress was verified through four measures, two were physiological in nature (continuous recording of galvanic skin response and pulse rate) and two were subjective (self-report scales of stress). The final database consisted of 48 speakers, 24 male and 24 female, all of whom produced deceptive statements while under a significant degree of stress, plus the high stress truth passages, and various low stress passages; additional speakers were also recorded to produce foil passages for balance.

The field-based studies consisted of a set of audiovideo recordings of military personnel answering questions while undergoing SERE training (Survival Escape Resistance Evasion). The SERE program is a rigorous survival training program where the students are trained not to reveal any information when captured and interrogated by hostile forces. The particular subjects took part in a guilty knowledge test in which they were instructed to lie about several aspects of SERE training. The goal of the study was to detect their lies from a large number of truthful responses. Since the subjects faced punishment if their lies were detected, they were lying under a substantial degree of

jeopardy. As with the laboratory materials, recordings from additional speakers were added as foils for balance.

The speech materials collected in the laboratory- and field-based studies were then used by three *evaluation groups* in assessing the accuracy of two VSA devices:

1. Two University of Florida's Institute for Advanced Study of Communication Processes (IASCP) team members who have received formal training and certification by the two VSA manufacturers, NITV and V.
2. Several trained representatives provided by each VSA manufacturer (three for CVSA; two for LVA); these groups were the most experienced in using the devices.
3. A third group, phoneticians who specialize in signal analysis (but without formal training by the manufacturer) evaluated one VSA device, CVSA. It was not appropriate to have the phonetician operator group evaluate LVA because the optimal method devised by the IASCP team to evaluate LVA did not require an operator. This group was the least experienced in using VSA devices.

Individual single-syllable words were sampled from the database for submission to the CVSA (this device requires such). Single sentences, along with a passage of calibration material, were submitted to the LVA (following the manufacturer's instructions). Both devices were tested separately for their sensitivity to deception and to stress. Sensitivity was defined in terms of the ability of a device to detect deception and stress when they are present (their "true positive" rate) and the frequency with which they incorrectly classify truthful/unstressed materials as deceptive/stressed ("false positive" rate). Sensitivity was also assessed by combining true and false positive rates into a single metric of sensitivity, d' ("d prime"), as well as other statistical tests.

The results from both the laboratory and field-based materials and from both devices were not positive. Both the CVSA and LVA performed at chance-level for deception, truth and stress. This result did not vary by operator class including the manufacturers' evaluation groups. True positive rates varied between 42% - 65% for the laboratory-based materials and 19% - 38% for the field-based materials. More importantly, false positive rates were high and very closely matched the corresponding true positive rates (40% - 70% for the laboratory-based materials and 41% - 49% for the field-based materials). These data appear to demonstrate that little to no relationship exists between the output of these devices and the actual presence or absence of deception and stress.

In considering the devices' performance in this study, an alternate interpretation might be considered. Specifically, the results may have indicated a failure to elicit sufficiently stressed/deceptive speech samples due to the inherent limitations of academic laboratory research. It could be argued that "real-world" levels of stress might be higher than the psychological stress which can be generated in a laboratory setting. This interpretation would be a difficult one to reject if only those speech samples that contained deception or stress had been examined. However, an assessment of the devices' performance on truthful and unstressed speech served as a control, one that permitted the examination of the device's potential bias to flag speech samples as deceptive in either the presence or absence of stress due to deception. If the database, collected under highly-

controlled conditions within the laboratory, was devoid of “real-world” levels of stress or deception, then very low false positive rates would be expected. Such was not the case.

In conclusion, a substantial amount of research has been completed over the course of this contract. First, a large database of highly controlled samples of deceptive and stressed speech has been established; so has another smaller one of several types of field materials. Second, personnel have been trained and certified in the use of two VSA systems. Finally, an evaluation has been carried out that is both basic and extensive.