AFRL-IF-RS-TM-2001-7
In-House Technical Memorandum
November 2001

INVESTIGATION AND EVALUATION OF VOICE STRESS ANALYSIS TECHNOLOGY

Darren Haddad, Sharon Walter, Roy Ratley and Megan Smith

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

20020610 039

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TM-2001-7 has been reviewed and is approved for publication.

APPROVED:

GERALD C. NETHERCOTT
Chief, Multi-Sensor Exploitation Branch
Info and Intel Exploitation Division
Information Directorate

FOR THE DIRECTOR:

JOSEPH CAMERA
Chief, Information & Intelligence
Exploitation Division
Information Directorate

If your address has changed or if you wish to be removed from the Air Force Research Laboratory Rome Research Site mailing list, or if the addressee is no longer employed by your organization, please notify AFRL/IFEC, 32 Brooks Road, Rome, NY 13441-4114. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
NOVEMBER 2001

3. REPORT TYPE AND DATES COVERED
In-House August 1998 - July 2001

4. TITLE AND SUBTITLE
INVESTIGATION AND EVALUATION OF VOICE STRESS ANALYSIS TECHNOLOGIES

5. FUNDING NUMBERS
C: N/A
PE: N/A
PR: NIJR
TA: SA
WU: 13

6. AUTHOR(S)
Darren Haddad and Sharon Walter
Roy Ratley and Meagan Smith

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
AFRL/IFEC
32 Brooks Road
Rome NY 13441

ACS Defense
P.O. Box 1188
Rome NY 13442

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFRL/IFEC
32 BROOKS ROAD
ROME NY 13441-4114

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TM-2001-7

11. SUPPLEMENTARY NOTES
AFRL/IFEC Project Engineer Darren M. Haddad, 330-2906

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Numerous police officers and agencies have been approached in recent years by vendors touting computer-based systems capable of measuring stress in a person's voice as an indicator of deception. These systems are advertised as being cheaper, easier to use, less invasive in use, and less constrained in their operation than polygraph technology. They claim that a speaker's medical condition, age, or consumption of drugs does not affect use of their system. Voice stress analysis does not require physical attachment of the system to the speaker's body and does not require that answers be restricted to "yes" and "no". Purportedly, according to some vendors, any spoken word or even a groan, whether recorded, videotaped, or spoken in person, with or without the speaker's knowledge, are acceptable inputs to voice stress analysis systems.

The value of voice stress analysis technology for military application could be extensive. During military field interrogations of potential informants, it could be applied in a manner similar to its application for law enforcement. Also, it's not known if stressed speech has any effects on the accuracy of speech technology, such as speaker identification and language identification. If voice stress can be detected, perhaps it can be taken into account in applying voice recognition technology and be used to improve these recognition capabilities. Therefore, this effort is to determine the scientific value and utility of existing, commercial voice stress analysis technology for law enforcement and military applications.

14. SUBJECT TERMS
Voice Stress, Speaker Identification

15. NUMBER OF PAGES
118

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLAS	UNCLAS	UNCLAS	UL

Table of Contents

1.0 EXECUTIVE SUMMARY 1

2.0 EFFORT OBJECTIVE 1

3.0 INTRODUCTION 1

4.0 HISTORY OF VSA TECHNOLOGY 2

5.0 AVAILABLE VSA SYSTEMS 5

5.1 PSYCHOLOGICAL STRESS EVALUATOR (PSE) 5
5.2 LANTERN 5
5.3 VERICATOR 5
5.4 COMPUTERIZED VOICE STRESS ANALYZER (CVSA) 6
5.5 VSA MARK 1000 6
5.6 VSA-15 6
5.7 XANDI ELECTRONICS 6
5.8 TVSA3 7

6.0 METHODS OF VOICE STRESS ANALYSIS AND CLASSIFICATIONS 7

7.0 TESTING 8

7.1 TEST OBJECTIVE 8
7.2 SCOPE/APPROACH 8
7.3 TEST AND ANALYSIS PROCEDURES 9
7.4 SYSTEMS TESTED 9
7.5 TECHNICAL TESTING 9
7.5.1 Artificial Signal Test (Test 1) *9*
7.5.1.1 Objective (Test 1) 9
7.5.1.2 Test 1 Set-Up 9
7.5.1.3 Vericator 10
7.5.1.4 Diogenes Lantern 10
7.5.1.5 Summary (Test 1) 10
7.5.2 Source Consistency Tests (test 2) *13*
7.5.2.1 Objective (test 2) 13
7.5.2.2 Scope 13
7.5.2.3 Test 13
7.5.2.4 Examination Results 14
7.5.2.5 Summary Test 2 15
7.5.3 Objective (Test 3) *15*
7.5.3.1 Data Evaluation (Test 3) 15
7.5.3.2 Data collection and Down Sampling 16
7.5.3.3 Segmentation 16
7.5.3.4 Testing 16
7.5.3.5 Results 16
7.6 FIELD TESTING 16

8.0 CONCLUSION 19

9.0 SUGGESTED FOLLOW ON 19

REFERENCES 20
Appendix A 21
Appendix B 22
Appendix C 25

List of Figures

FIGURE 1: FM RECORDER TEST SIGNALS @ 80 HZ & 160 HZ 10
FIGURE 2: WAVERFORMS TO THE DIOGENES LANTERN SYSTEM 12
FIGURE 3: TEST CONFIGURATION FOR SOURCE CONSISTENCY TESTS 13
FIGURE 4: WAVEFORM CHANGES USING THE CASSETTE RECORDER WITHOUT THE AGC SET 15

List of Tables

TABLE 1: COST COMPARISONS MADE BY A VSA VENDOR 2

Preface

Relationship between AF and NIJ:

The unique relationship between the Air Force, Rome Research Site and the National Institute of Justice (NIJ), was established over four years ago with a Memorandum Of Understanding (MOU) agreement. This agreement allowed the Air Force the opportunity to test and evaluate their technologies for law enforcement applications. With this agreement, the Air Force was able to collect a variety of law enforcement audio/video data that would eventually be utilized to test the performance of Air Force algorithms developed within the Rome Research site at Rome, New York. This enables the NIJ to receive the capability to demonstrate technologies that may apply to law enforcement encouraging the possibility transfer of those technologies to state and local crime fighting units.

1.0 EXECUTIVE SUMMARY

Voice Stress Analysis (VSA) systems are marketed as computer-based systems capable of measuring stress in a person's voice as an indicator of deception. They are advertised as being less expensive, easier to use, less invasive in use, and less constrained in their operation than polygraph technology. Law enforcement officials have inquired about this technology. As a result, the National Institute of Justice (NIJ) has petitioned the Air Force Research Laboratory (AFRL/IFE) for assistance in evaluating voice stress analysis technology. This evaluation is broken down in three phases. In the first phase, Dr. John H.L. Hansen, from the University of Colorado, investigated the feasibility of detecting stress from speech. He reported on the methods, analysis, and classification of voice stress contained in the appendix of this report. The second and third phase of this study investigated the reliability of commercial VSA units, from a theoretical point of view and from an application (i.e. law enforcement) point of view.

2.0 EFFORT OBJECTIVE

The Objective of this effort is to determine the effectiveness of commercially available voice stress analyzers (VSA) to detect "stress" in the voice of a talker. The use of "stressed speech" for this effort is defined as speech that exhibits a change in characteristics caused by mental stress such as anxiety and/or fear. Of particular interest is the detection of stressed speech (change) caused by an act of deception under law enforcement interview questioning or military interrogation.

3.0 INTRODUCTION

Police departments everywhere are bombarded with offers of advanced technologies by commercial enterprises that promise to reduce their officers' workload, improve law enforcement effectiveness, and/or save lives. With increasingly limited budgets, police departments must turn a critical eye to every purchase.

One interest by law enforcement and military organizations are the commercial VSA systems, which are advertised to detect deception or to detect when a person under interrogation is lying. If voice stress can be detected, and effectively analyzed, perhaps it can be used as a viable investigative tool as well as an adjunct to speech recognition technology in order to improve speech recognition capabilities.

Numerous police officers and agencies have been approached in recent years by vendors touting computer-based systems capable of measuring stress in a person's voice as an indicator of deception. These systems are advertised as being cheaper, easier to use, less invasive in use, and less constrained in their operation than polygraph technology. Table 1 is a replication of the table of comparisons made by one vendor contrasting their VSA system with a computerized polygraph. Besides costing less to purchase the equipment and train users, the table indicates that a VSA examiner can conduct seven (7) exams per day while a polygraph examiner can conduct only two (2) per day. This vendor claims to always have conclusive results, and the ability to analyze recorded audio as well as live

speakers. They claim that a speaker's medical condition, age, or consumption of drugs does not affect use of their system. Voice stress analysis does not require physical attachment of the system to the speaker's body and does not require that answers be restricted to "yes" and "no". Purportedly, according to some vendors, any spoken word or even a groan, whether recorded, videotaped, or spoken in person, with or without the speaker's knowledge, are acceptable inputs to voice stress analysis systems.

The value of voice stress analysis technology for military application could be extensive. During military field interrogations of potential informants, it could be applied in a manner similar to its application for law enforcement. Also, it is not known if stressed speech has any effects on the accuracy of speech technology, such as speaker identification and language identification. If voice stress can be detected, perhaps it can be taken into account in applying voice recognition technology and be used to improve these recognition capabilities. Therefore, this effort is to determine the scientific value and utility of existing, commercial voice stress analysis technology for law enforcement and military applications.

Table 1: Cost comparisons made by a VSA vendor

	Computer Voice Stress Analyzer	Computerized Polygraph
Initial cost of system	$9,250.00	$13,000.00
Tuition for 1 student	$1,215.00	$3,000.00
Length of training	6 days	8 weeks
Cost of room and board factored at $70 per day	$420.00	$3,920.00
Salary for student while in training (U.S. average)	$769.23	$6,153.84
Number of exams that an examiner can conduct per day	7 exams	2 exams
Average percent of inconclusive results on exams	0%	20%
Can unit analyze audio tapes for truth verification?	Yes	No
Do drugs, medical condition, or age affect testing?	No	Yes
Total cost to purchase 1 unit and train 1 agent	$11,654	$26,073.84

4.0 HISTORY OF VSA TECHNOLOGY

In 1970, and prior to the publishing of Lippold's article in 1971, three military officers retired from the U.S. Army and formed a company which they named Dektor Counterintelligence and Security (CIS). The three officers were Alan Bell, Bill Ford and Charles McQuiston. Bell's expertise was in counterintelligence, Ford's was in electronics, and McQuiston's was in polygraphy. Ford had invented an electronic device

that utilized the theory of Lippold, Halliday and Redfearn in which he tape-recorded the human voice, slowed it down three to four times its normal rate, and fed it through several lowpass filters which then fed the signal into an EKG strip chart recorder. The strip chart recorder then made chart tracings on heat sensitive paper. They named their device the Psychological Stress Evaluator (PSE). Although Dektor CIS was intended to be a security company, the PSE immediately became a success and their focus became centered on this system. One of the first individuals hired by Dektor was a polygraph examiner with a local police department which had started utilizing the PSE. This individual, along with McQuiston, wrote a three-day training course based on their polygraph experience and utilizing polygraph formats.

According to Allan Bell Enterprises [1], "All lie-detection examinations or evaluations are predicated upon the fact that telling a significant lie will produce some degree of psychological stress. Psychological stress, in turn, causes a number of physiological changes." Polygraph takes advantage of these physiological changes to measure one's psychological stress. Polygraphs customarily measure changes in blood pressure, hormone levels, stomach and chest breathing patterns, galvanic skin response (perspiration), the pulse wave and amplitude.

VSA literature [9] points to a descriptor of the physiological basis for the micro muscle tremor or microtremor. This paper describes "a slight oscillation at approximately 10 cycles per second" (i.e. physiological tremors) during the normal contraction of voluntary muscle. All muscles in the body, including the vocal chords, vibrate in the 8 to 12 Hz range. It is these microtremors that the VSA vendors claim to be the sole source of detecting if an individual is lying. This human system is a feedback loop, similar to a thermostat/heater that will maintain an average temperature. By raising the temperature a little above the setting, it will switch off, and not come back on until the temperature is a little below it. Just as the temperature swings up and down over time, so too do the muscles tighten and loosen as they seek to maintain a constant tension. In moments of stress, the body prepares for fight or flight by increasing the readiness of its muscles to spring into action. The muscle vibration increases. This muscle tremor is usually evident in a hand tremor, as when one holds their arm out in an extended position. This indicates that restricting the blood supply to the muscle can reduce the tremor. Physiological tremor is "the ripple that is superimposed on the voluntary contraction of a particular muscle and arises solely from this activity." Most people exhibit a fine, rapid tremor of their hands when their arms are outstretched. According to the Merck Manual [12], "enhanced physiologic tremor maybe produced by anxiety, stress, fatigue, or metabolic derangements (ie. alcohol withdrawal, thyrotoxicosis) or by certain drugs (ie. caffeine and other phosphodiesterase inhibitors, beta-adrenergic agonists, and adrenal corticosteroids beta-blocker: propranolol)."

The initial VSA development entitled "Application of Voice Analysis Method" was funded by the U.S. Army Land Warfare Laboratory, performed by Decision Control, Incorporated of Bethesda, Maryland. This study was performed to assess the capability of a method of voice analysis to detect stress in the spoken response, "no." The studies recorded voice responses of individuals undergoing polygraph testing and were analyzed for their stress values. The results were then compared to the polygraph interpretations.

In this stress response comparison, the waveform results were similar. A prototype voice analyzer was developed, fabricated and tested. The device processed recorded audio and provided three voice measures. The introduction to this report indicates that a previous study [14], had shown that an analysis of the response "no" could provide an accurate assessment of whether the response was truthful or deceitful. Six semi-orthogonal measures and a number of bandpass frequencies were used in the study. The experiment simultaneously used the polygraph to determine the existence of stress. The results concluded that it was highly desirable to reduce the number of measures, and to determine the best set of bandpass frequencies.

U.S. Army Land Warfare Laboratory Technical Report #LWL-CR-03B70 by Joseph F. Kubis of Fordham University, titled "Comparison of Voice Analysis and Polygraph as Lie Detection Procedures" (commonly referred to as "the Kubis Report."), completed a study comparing the two types of lie detecting systems [8]. Two voice analysis systems were evaluated as lie detection devices in a simulated theft experiment, which utilized 174 subjects. One group of subjects was examined with the polygraph, at the same time their voice recordings were taken. A smaller group was tested only with their voice being recorded. The results failed to demonstrate that either of the voice analysis systems were accurate in identifying the three basic roles of Thief, Lookout, and Innocent Subject in a simulated theft experiment. The polygraph achieved an accuracy score of 76 percent, a value comparable to that obtained in previous studies using the simulated theft paradigm. Independent raters, who knew nothing about the characteristics of the experiment subjects, also obtained 50 to 60 percent accuracy scores in the examination of the polygraph charts. In the Kubis report, the results showed that the voice recordings were not statistically significant. It showed that lower accuracy using voice analysis was obtained with voice recorded and polygraph-tested subjects than with those who had their voice recorded without the polygraph. Audio recording monitors that were present during the interrogation sessions based their judgements more on their perceived impressions of the suspect rather than the output of the system. They were able to discriminate among Thief, Lookout, and Innocent Suspect. Based on these results, one could hypothesized that the simulated theft procedure induced a sufficient degree of emotional stress on a subject which indicates that is could be useful for lie detection research.

Another study [2], takes issue with the Kubis Report, citing the experimental methodology as a "game model" with possibly insufficient induced stress for measurement, a noisy environment, and deviations from manufacturer-recommended questioning techniques.

Polygraph-licensing laws in some states require lie detection to be accomplished specifically with a polygraph. These laws define lie detection equipment as equipment providing a permanent record of cardiograph (heart) and pneumograph (breathing) data. The earliest law was enacted in Kentucky in 1962. These legislation enactments made the VSA units illegal in some states. State-sponsored hearings in Florida in 1973 and 1974 resulted in an informal acceptance of the PSE for law enforcement use in that state. North Carolina and Arkansas soon followed and formally authorized use of the PSE.

5.0 AVAILABLE VSA SYSTEMS

Currently, there are many available VSA on the market today. The major VSA vendors market their products on a laptop with specific software, while a few are sold as an electronic device with the software embedded on its chips. Examples of VSAs currently available are described below.

5.1 Psychological Stress Evaluator (PSE)

Dektor Counterintelligence and Security, Inc. of Springfield, Virginia. The Canadian Patent #943230 (March 5, 1974) and United States Patent #3,971,034 (July 20, 1976), submitted by Allan D. Bell, Jr., Wilson H. Ford, and Charles R. McQuiston, describe their "Physiological Response Analysis Method and Apparatus."

This unit was the first VSA unit on the market, released in March of 1971. It was designed to be used in the same manner as a polygraph, one-on-one testing for the detection of deception. It was a black box with an output in the form of a waveform via thermograph readout. The PSE senses the difference and records the change in the inaudible FM qualities of the voice on a chart. When an experienced examiner interprets the chart, it reveals the key stress areas of the person being questioned.

5.2 Lantern

The Diogenes Group, Inc.
(407) 933-4839
FAX: (407) 935-0911

The Diogenes Group Inc., established in 1995, produces a system called the Lantern. The Lantern instrumentation consists of an analog-type magnetic tape recorder with integral microphone, a Pentium laptop computer serving as a high-speed processor, and an extensive program of copyrighted, proprietary processing software designed specifically for ease of operation. The Windows 3.11™ or Windows 95™ based software is also responsible for control of all processing operations, display format and presentation, and the printing of hard copies of the waveforms representing the behavior of the microtremor. The tape recorder is operated throughout an interview to create the primary record, which includes both questions and answers in the context in which they occurred. The monitor output of the recorder provides the real-time input to the digital processor. The examiner is able to control, with a single finger, high-sample rate digital capture of the sound of each answer. [6]

5.3 Vericator

Trustech Ltd.
Integritek Systems, Inc.,
111 Bermuda Ave.
Tampa, FL 33606
+1 813 250 3922

Trustech Ltd. was founded in 1997, and produces a system called Vericator, formally known as the Truster Pro. This system allows the user to use their own personal computer with the following requirements: WIN95 ™ / WIN98 ™ /NT 4.0 ™, Pentium ™ II or III, 32 MB RAM to 128 MB RAM, a microphone, CD ROM Drive (double speed), and a16 Bit Soundcard (full duplex). The package includes a Vericator CD , Stereo T-Connector (for connecting your PC and telephone), Vericator User Manual. It features automatic calibration process; analysis of deception in real-time; analysis of pre-recorded online conversations/interviews and TV or radio segments. The summary and technical reports can be viewed, saved and printed. There are graph displays for advanced diagnosis; four built-in psychological lie detection patterns; filtering system for reducing background noise. [7]

5.4 Computerized Voice Stress Analyzer (CVSA)

National Institute for Truth Verification (NITV)
West Palm Beach, Florida 33414
(561) 798-6280
FAX: (561) 798-1594

In 1986, NITV began to market the Computerized Voice Stress Analyzer (CVSA), currently known as the most popular VSA system available. NITV advertisements claims that the system is in use in more than 500 law enforcement agencies, and offers as evidence, letters of endorsements from agencies throughout the United States. NITV claims to market only to law enforcement agencies in order to prevent it from being used by criminals to identify undercover agents. [5]

5.5 VSA Mark 1000

- Marketed by CCS International, Inc.:

This system is marketed as a covert electronic lie detection system providing fast analysis, fast results and fast answers. With its built-in tape recorder, the VSA Mark 1000 allows you to analyze audio data at a later time. A clear, precise digital readout is given in both LED and printed format, where the results are instantaneous. For more information go to http://www.spyzone.com/catalog/index.html. [4]

5.6 VSA-15

- Marketed by CCS International, Inc.:

This system is similar to the VSA Mark 1000, but is marketed as a miniaturized hand held system. This unit is targeted for the non-professional user. For more information go to http://www.spyzone.com/catalog/index.html. [4]

5.7 Xandi Electronics

- Markets a Voice Stress Analyzer Kit (Model # XVA250) for $59.

It has 10 LEDs. The System is powered by either a 9 volt battery or a power adapter. As you speak into the analyzer, the LEDs in the normal position (the left) should light up.

Under stress conditions, more of the LEDs on the right-hand side will light up in the stress position, and fewer will light up in the normal position. [11]

5.8 TVSA3

- This VSA software is freeware off the World Wide.

The TVSA3 is a software program, which inputs digital audio files, and outputs new audio files mixed with a changing tone in the background. These background tones indicate the changing stress levels of the individual that is speaking. A higher tone indicates a higher stress level. The lone control is a threshold setting, which determines how high the voice stress frequency must be to trigger the background tone. The threshold setting is treated as a percentage of the stress range found in a given recording. [3]

Only the Diogenes-Lantern and Vericator were assessed in this study and will be discussed in this report. These are the most popular VSA units available on the commercial market today. Another popular unit is the CVSA, but the company decided not to participate in this study.

6.0 METHODS OF VOICE STRESS ANALYSIS AND CLASSIFICATIONS

To better understand the aspects of stress speech in a human, the Air Force Research Laboratory (AFRL) worked with Dr. John Hansen of the University of Colorado, to determine if it is feasible to recognize and classify stress in an individual's voice. Dr. Hansen is a world known expert in the area of voice stress. The report is included in this report, and is attached as an Appendix C. He states " it is not inconceivable that under extreme levels of stress, that muscle control throughout the speaker will be affected, including muscles associated with speech production". In this study he used the Speech under Simulated and Actual Stress (SUSAS) database. This database includes stress speech such as angry, loud, lombard (speaking under noisy conditions), and fear stress. In his report, he reviewed literature that discussed past speech under stress studies. He analyzed stress in speech, in which he concluded that voice stress is caused by factors that introduce variability into the speech production process. These variabilities or features include duration, glottal source factors, pitch distribution, spectral structure and intensity. Duration includes four area: (1) overall word duration, (2) individual speech class (vowel, consonant, semivowel, and diphthong) duration, (3) duration shifts between classes, and (4) speech class duration ratios. Glottal source factors measured the spectral slope of those vowels, which were longer than 5 frames or 96 msec. The first and second formants locations are measured to determine the spectral structure. Intensity is a calculation of energy in an voice signal. These variabilities could also be speaker dependent. By using these various linear and nonlinear features, and testing with the Bayesian hypothesis method, it was concluded that different types of emotional stress could be classified. The Bayesian hypothesis method is a stress detection technique to determine if a given piece of audio data is either neutral speech or a certain classification of stress speech. From the results, it suggests that it is unlikely that a single feature could be used to accurately detect deceptive stressed speech. The more features that are fused

together, the stress type recognition improves. It also shows that some features, single handed, can detect a specific type of stress better than other features. For an example, the pitch feature could detect loud stress better than angry and lombard stress. Whereas, the spectral structure feature could detect angry stress better. Classification of deceptive stress was not tested due to the unavailability of a deceptive database. The collection of a deceptive database is a recommendation of future work (see section 8.0).

7.0 TESTING

The goal of these tests is to determine how effective these VSA units can detect stress. VSA vendors have marketed their technology as scientific, as it takes advantage of the human micromuscle tremor in the vocal tract. These tests attempt to prove or disapprove these theories.

7.1 Test Objective

The objective of these tests is to measure the output response of two VSA systems, given several controlled input signals. This will be used to verify the manufacturer's claims of operation for each analyzer. The degree of source consistency of results for each analyzer will then be determined. This will determine the correct process to use when recording audio for evaluation. Finally, the VSA systems will be laboratory tested and field tested by evaluating them with trained laboratory analysis and experienced police investigators.

7.2 Scope/Approach

This effort will test and evaluate two (2) commercially available voice stress analyzers. Tests will be accomplished using a series of test signals that contain information distributed over the frequency spectrum, generally covered by the spectrum of normal speech. Analysis of the VSA test results will be conducted to determine

- VSA response characteristics
- Degree of accuracy compared to the manufacturers theory of operation and technical specifications
- Accuracy of result repeatability
- Evaluate under real world conditions

The test approach taken in this plan is to consider each analyzer to be a black box, and to record its output response to known input test signals.

7.3 Test and Analysis Procedures

The procedures are developed for three areas - procedures for the development of test tapes containing artificial signals, source consistence test, and analysis and evaluation of audio data with stress ground truth.

7.4 Systems Tested

Vericator and Diogenes Lantern

7.5 Technical Testing

Trained analysis in a laboratory setting completed the technical testing. These analysis were each trained through the VSA vendor training programs.

7.5.1 Artificial Signal Test (Test 1)

7.5.1.1 Objective (Test 1)

Test 1 of the VSA Evaluation was to determine if the VSA units detect the frequency modulation of a signal. These signals are similar to the microtremor, which manufacturers state is their theory of operation. For the purposes of this test we utilized the Vericator system and the Diogenes Lantern system. A generically generated signal database of FM frequencies, occurring at different rates and depths of modulation, was processed repeatedly through the systems.

7.5.1.2 Test 1 Set-Up

The test was performed on laptop computers that contained the Vericator and Diogenes Lantern software. The signals were fed to the laptops from a desktop PC. The desktop PC dispatched the artificial signals using the commercial off-the-shelf (COTS) application Cool Edit. Cool Edit is a digital audio editor for a Windows base system. It is used to record and play files in a wide variety of audio formats, edit files and mix them together, and convert audio files from one format to another. Cool Edit also gives the ability to create sounds from scratch with generated tones and generated noise signals.

The FM test signals that comprise the signal database for Test 1 were generated using Cool Edit. These FM signals were generated at the carrier frequencies of 80 Hz and 160 Hz (these frequencies represented the fundamental frequency on a speech signal), with varying modulation rates and depth of modulation rates (Figure 1). Modulating rates measures how fast the signal modulates, and depth represents how much the signal modulates from the carrier frequency.

DEPTH OF MODULATION (Hz)	MODULATION RATE (Hz)							
	1	2	4	6	10	15	20	25
1	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X
4	X		X	X	X	X	X	X
6	X			X	X	X	X	X
8	X				X	X	X	X
10	X					X	X	X
15	X						X	X
20	X							X
25	X							X

Figure 1: FM Recorder Test Signals @ 80 Hz & 160 Hz

The test signals were recorded in 15 second utterances. Each signal was passed through each VSA system. The test results were recorded on data spreadsheets, and the wave analysis was labeled and printed. Once the test waves were all analyzed the documentation was compared to determine consistency.

7.5.1.3 Vericator

For the purposes of this test we utilized the "Online Mode" of the Vericator application. The "Online Mode" measures five voice parameters SPT, SPJ, JQ, AVJ, SPLC-SOS (see appendix A) in real time (< 2 second delay) to detect stress. The signals were processed through the Vericator in short utterances. A few signals were attempted with a consistent result of "No indication of voice segments" or "Not enough voice samples."

To overcome the inability to analyze the bare FM tones, we added a voice to the signal to the tone. After recording a female voice, analyzing it and determining the most consistent signal, the FM frequency was added. The system was able to identify the signal and process it. The system responded with a spike in the analysis wave every time the FM frequency was introduced to the signal. These results were recorded in the test log for the Vericator analysis. The signals listed in figure 1 were processed through the Vericator system.

7.5.1.4 Diogenes Lantern

The test waves were re-sampled to an 11025-sampling rate, 8-bit mono to facilitate the acceptance of the signal by the Diogenes Lantern system. The FM signals were 3-4 seconds long. The signals were processed through the Lantern system. The graphs were labeled according to frequency, depth of modulation, and modulation rate. The signals listed in figure 1 were processed through the Diogenes Lantern.

7.5.1.5 Summary (Test 1)

The tests were performed, the data was documented, and the results were compared. The Vericator and Diogenes Lantern Systems were utilized in this evaluation and their technology was tested. The primary goal of this phase of the VSA evaluation was to determine if the microtremor claim is the VSA's true theory of operation. For the purposes of this test the nature of the results, stress or no stress indicated, were not taken into account. The results

were found to be consistent across the board with little variation in the results in response to the adjustments/changes in the modulation or depth of modulation rates. For example, the analysis of the 80Hz FM test wave, with a depth of modulation rate of 1 Hz and a modulation rate of 1 Hz, differed very little from an 80 Hz FM test wave with a depth of modulation of 4 Hz and a modulation rate of 25 Hz. Since there was no variation of indicated stress from different input signals, it can be assumed that the systems tested do not use microtremors as indicated in their claims.

It was determined, late in the testing phase of this project, that the Diogenes Lantern System measure the energy change of the spectrum envelope between 20 Hz and 40 Hz. This is what the Diogenes Lantern System claims to be microtremors. It is the change of energy in the speech envelope. If an individual is under stress, their vocal tract muscles are likely to tighten up. When the vocal tract muscles tighten up the energy of the voice signal becomes abrupt when the individual starts and finishes talking. During the time an individual talks, there is less variation of energy within this the 20 Hz bandpass. When an individual is not stressed, their voice energy slowly leads to a peak when they start to speak, then the energy varies until the individual stops speaking where the energy slowly tails off. This algorithm was coded in the laboratory with the same audio signal inputted. As seen in the waveforms in figure 2a and 2b, the results were identical when compared to the Diogenes Lantern system. The waveform comparison could also be seen in figure 2c and 2d. These figures prove that the algorithm used in the Diogenes Lantern system is energy based. This discovery makes the artificial signal test obsolete since the objective was to determine if these units detect frequency modulation of an audio signal.

The Vericator claims that it analyzes multiple features of speech to determine if an individual is lying. It is was not proven if this claim is true, since this information is proprietary, nor was it proven what speech features are being analyzed. However, it is likely that they do process multiple algorithms simultaneously due to the multiple waveforms being display. Since, the Vericator did not react to this test, it is safe to say that the measurement of micromuscle tremor is not one of the speech feature algorithms being used in their system.

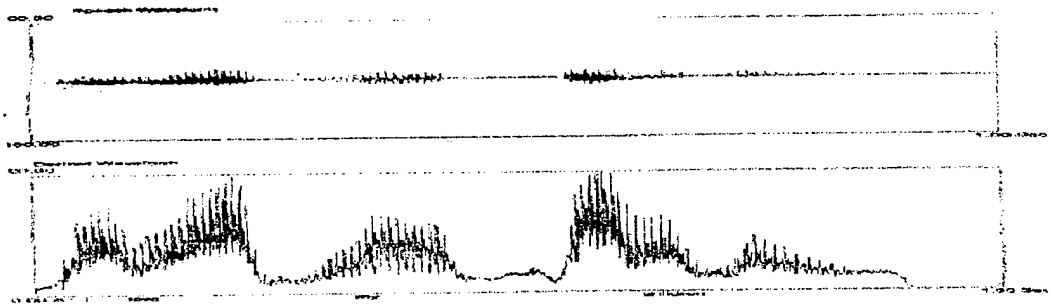


Figure 2a Diogenes Lantern System Output (No Stress Indicated)

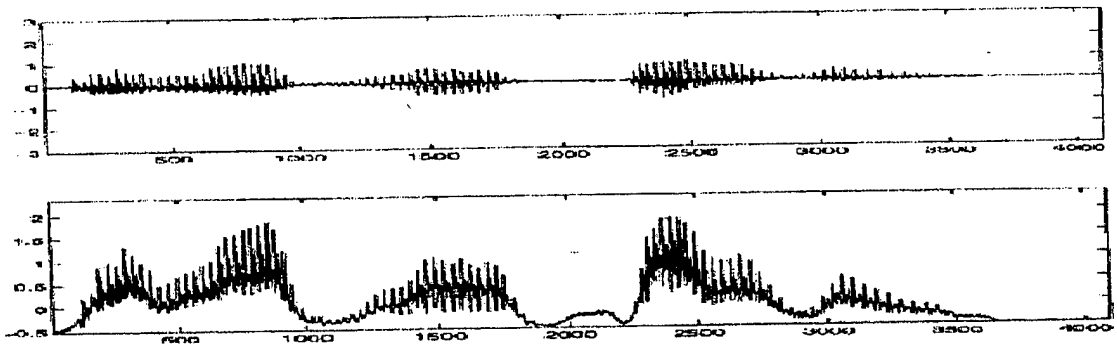


Figure 2b Matlab Output (No Stress Indicated)

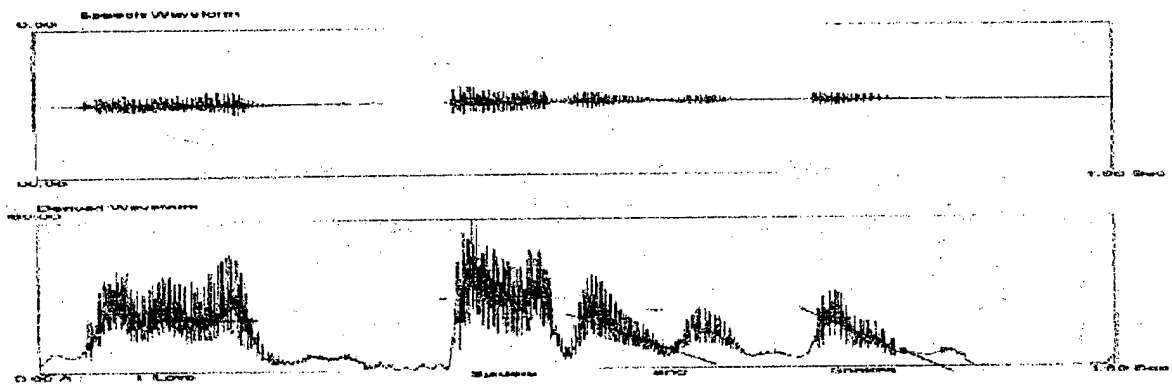


Figure 2c Diogenes Lantern System Output (Stress Indicated)

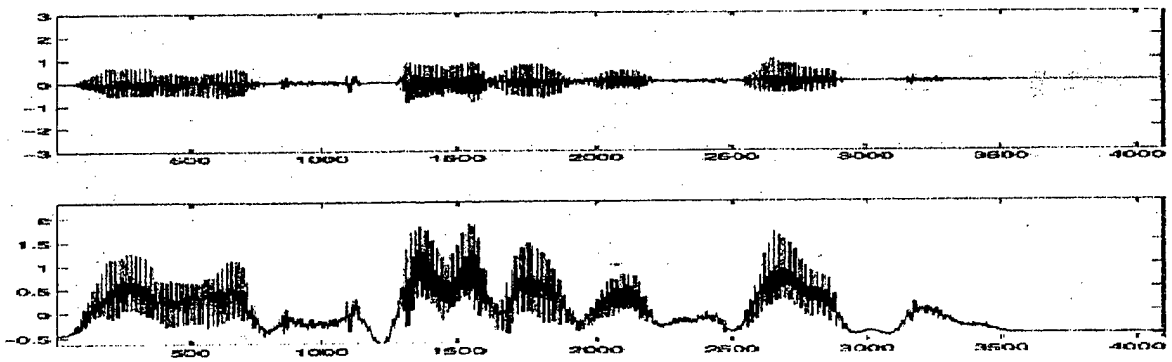


Figure 2: Waverforms to the Diogenes Lantern system

7.5.2 Source Consistency Tests (test 2)

One of the major questions presented to the engineers testing the voice stress systems, "is there a difference in the analysis of an audio file utilizing different medias?" The different medias could be a Digital Audio Tape system (DAT), a cassette recorder, or telephone input device. Each recording device has their own different properties, which could effect the overall analysis by the examiners.

7.5.2.1 Objective (test 2)

This experiment is designed to compare the analysis of identical signals utilizing the different medias.

7.5.2.2 Scope

Identical signals were fed several times into these medias, according to figure 3, to evaluate the consistency of the results from the two VSA systems. The analysis of the output was then compared to the analysis of the output of the same signal from a different type of media. This gave indications of whether or not different types of medias play an important role in the evaluation and analysis of the voiced responses.

7.5.2.3 Test

AFRL and ACS Defense jointly collected 60 voiced utterances from different males and females and recorded those voice utterances on DAT, computer and cassette media. These utterances were collected simultaneously by the computer (. wav format), analog cassette format, and digital via a 48KHz Digital Audio Tape (DAT) recorder (see figure 3). The audio was analyzed separately from each of the three medias (cassette, computer, and DAT). The live feed was connected directly into the VSA computer, the output was analyzed and the results were printed. At the same time, the utterance was recorded on the DAT, this signal was latter processed in the VSA unit for reanalysis and the results were printed. Again at the same time, the utterance was recorded on a cassette tape and this signal was latter processed in the VSA unit to be re-analyzed, and again the results were printed.

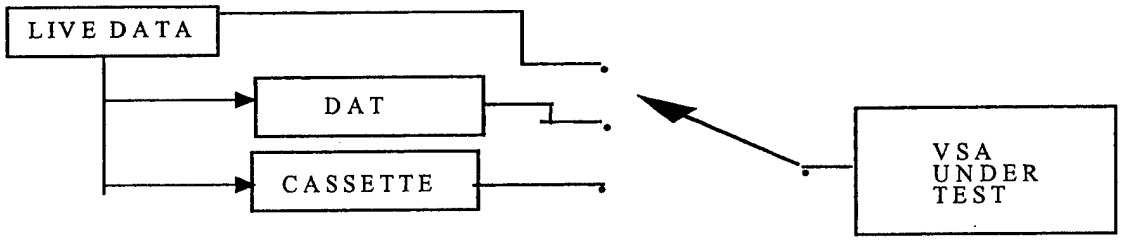


Figure 3: Test Configuration for Source Consistency Tests

7.5.2.4 Examination Results

Voiced analysis reported consistent results utilizing DAT and live voice. Each utterance was examined and found that all the waveforms and analysis was consistently identical. When using a cassette recorder similar results were obtained as in the live data. When recording with a cassette player, care needs to be taken when adjusting the automatic gain control (AGC). If the recording volume is not set accurately, the input voice signal gets clipped, so when the output waveform is processed it gets distorted too. This could result in an analysis which is completely different from the truth, therefore providing and incorrect result by the examiner. These discrepancies can be seen in figures 4a - 4d, when using the Diogenes Lantern system. When using the Vericator these discrepancies are also evident. The Vericator results reacted differently each time the same clipped data was inputted into the system. For example, if a clipped audio segment was processed in the Vericator, the system may display truth, while at another time that same clipped data would cause the system to display false statement.

Reviewing the charts in figure 4, shows how much the waveform will change when recording with the cassette recorder without the AGC set. The input file (top waveform) is consistent for figure 4a, 4b, and 4c. It is clipped for figure 4d. This corrupts the output signal (bottom waveform), as seen when comparing figure 4d with the others. Other waveforms can be reviewed in appendix B.

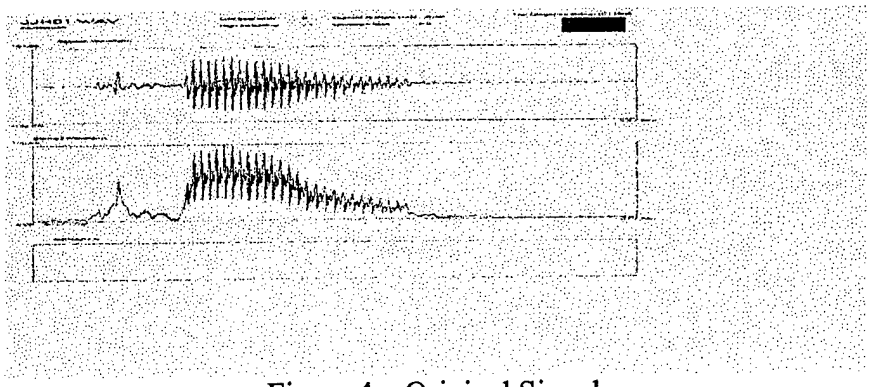


Figure 4a: Original Signal

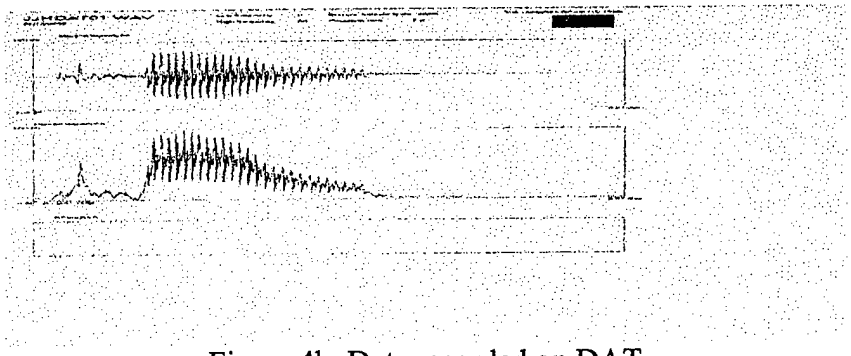


Figure 4b: Data recorded on DAT

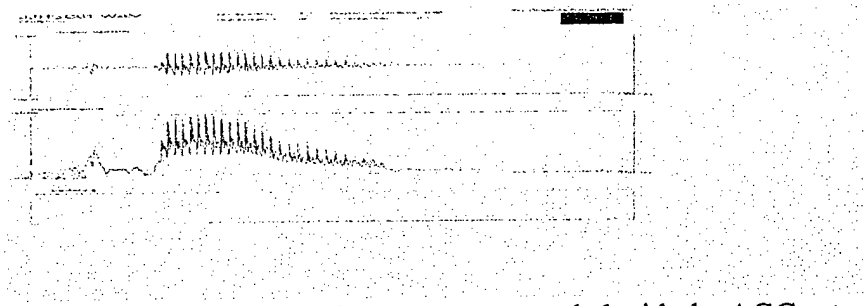


Figure 4c: Data recorded on cassette recorded with the AGC set

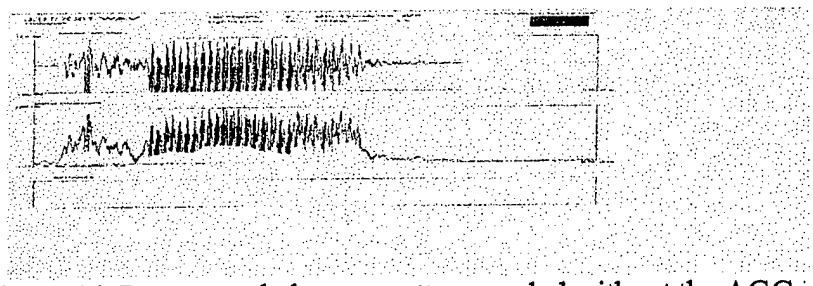


Figure 4d: Data recorded on cassette recorded without the AGC set

Figure 4: Waveform changes using the cassette recorder without the AGC set

7.5.2.5 Summary Test 2

From the results in test 2 it is recommended to perform all analysis, when recording, using the DAT as a recording media. This eliminates any media effects on the audio signal and provides consistent results. It is also absolutely necessary to use a Shure microphone model SM58, or one with equivalent specifications. When using the cassette recorder, it was shown that human error could change the results of any findings. As shown, this type of recording distorts the input audio signal, therefore providing the VSA units clipped audio data to process.

7.5.3 Objective (Test 3)

The next stage of the VSA evaluation, consisted of assessing the results of known-ground-truth data when processed through the Vericator and Diogenes systems. Segmented audio data was administered to the systems from law enforcement cases that have been solved.

7.5.3.1 Data Evaluation (Test 3)

Data: Audio statements from 2 sets of polygraph tests performed by a certified polygraphist

Evaluators: By analysis who were certified by Diogenes and Vericator manufacturers.

7.5.3.2 Data collection and Down Sampling

Six videotapes were obtained from DODPI, of two suspects in two separate murder cases. The audio portion of the videotape was extracted and digitized into .wav files. The audio files were then down sample from 48kHz down to 11.025kHz to accommodate the manufacture's requirements. This process is necessary to make the data compatible to the VSA units, since these units are programmed to accept data at the 11.025 kHz sampling rate. These digital audio files were inserted into the VSA computers.

7.5.3.3 Segmentation

Once the exact audio data was entered and stored in the two computer systems we then proceeded to segment the audio. For the Lantern system we had to create individual .wav files for each utterance that the defendant made, usually answered by a yes or no in these cases. This was done to allow short utterances to be processed by the Lantern as suggested by the manufacturer. There were a total of 45 questions ranging from relevant to non-relevant questions.

The Vericator performs it's own unique segmentation. We segmented the audio utilizing their own process. This was done through the off-line mode.

7.5.3.4 Testing

Each audio segment was processed through the Lantern system and performed a separate analysis of each wave pattern. Each waveform was compared to the other to verify any distinct changes due to stress. Each file that gave indication of stress were marked and compared to the baseline.

Each audio file was processed through the off-line mode of the Vericator. Results were automatically recorded by the system.

7.5.3.5 Results

The stress ground truth was obtained through the polygraph examiner and court proceedings via the outcomes of each of the interviews. Both suspects confessed and were subsequently convicted of murder. All of the relevant stress sentences were verified. Each of the 48 utterances was analyzed and compared to the ground truth. Each system gave indications of high levels of stress where stress indicators were verified. The Vericator system scored 100% in its indication of some form of stress, where as it displayed deceitful, high stress, or probably lying. The Lantern system also scored 100% in its indication of stress through the waveform analysis. Both systems gave the examiner a conclusive indication of relevant stress.

7.6 FIELD TESTING

In the field testing portion of this study, two local police investigators obtained a VSA system, Mike Adist of the Canastota, New York Police Department and James F. Masucci of the Rome, New York Police Department. Mike Adist used the Vericator and James Masucci

used the Diogenes Lantern. The goal of this phase of testing was to determine the feasibility of these systems in the law enforcement environment. It also provided the unbias opinion of an experienced investigator. The following are their reports:

I have been in Law enforcement for the part twenty-one years, and during this time I have had the opportunity to see all facets of crime and investigations. I have been involved in crimes dealing with the least punishable to the severest of them all. I have had the opportunity to attend schools that taught me how to detect when a suspect is being deceptive during questioning. In some cases it was difficult to determine if a suspect was deceptive, and that made my job harder until the summer of 1997 when I came to the Law Enforcement Analysis Facility (LEAF) for help.

My first contact was with Sharon Walters who advised me that the U.S. Government (Military) and a group of Research Technicians (Private Contractors) at the Rome Research and Technology Facility were about to take on the task of evaluating some technology dealing with truth verification. I was also told that this evaluation might be effective in Law Enforcement. At that time I was pleased for many reasons.

I was asked to join this task force to assist the government in this evaluation, but first I was to learn what truth verification was. This required me to learn and study what a microtremor was, and how algorithms mathematically calculated the stress in a human voice. I reviewed the technology and was given a voice stress program called Truster-Pro, now known at Vericator. Using this system I was able to interview a subject who may have been involved in a crime. First an interview is performed to determine the facts, as he/she knew them. Then, I was able to give the subject one or two tests to determine the truth or deception. Finally, a post interrogation would be conducted in an attempt to get a confession.

Keeping the voice stress technology in mind during the testing of a subject, one should remember that this type technology in it's self is only as an investigative tool, and cannot be use to convict the subject. Along with observing a subject's involuntary movements such as his eyes, legs and hands I have had great success the voice stress technology. I have had the opportunity to use this technology on crime from Petty Larceny to Rapes, and have been able to determine either from the victim (s) or the suspect (s), the deception or truth. Not all of the testing were positive, but on the majority of them I was able to get true confessions to the crime. Over the past three years I would say that I have achieved a success rate of about 97 percent on tests vs. confessions. I believe in this systems capability of becoming a valuable investigative tool for the law enforcement officer on the streets of our cities, towns and villages across the nation.

Respectfully,

Michael G. Adsit
Criminal Investigator

I have been using the Lantern Voice Stress Analyzer from Diogenes since October of 1997. I have had many rewarding experiences with the Lantern. I have successfully used it in homicide, arson, robbery, burglary, assault and sexual abuse cases. I do all of the testing for the Oneida County Child Advocacy Center, formerly known as the Oneida County Sexual

Abuse Task Force. I point this out to show that I have tried the Lantern on just about every type of crime. Although I did not keep statistics, I feel that I can safely say that with the aid of the Lantern, I have been able to eliminate about as many suspects as I have found reason to "dig into " a little more.

I am not much of a technical expert, but I have made the following observations. I do believe the theory of the micro-muscle tremor and the need for "jeopardy." I have found that without jeopardy, or a fear of some consequence to lying, you do not get accurate charts. I have seen a tremendous difference in the voice stress patterns when there is jeopardy – vs – no jeopardy. For example, I have told suspects to intentionally lie on certain questions during the test. I have found that when they do lie over something that means nothing, you don't get a clear-cut stress pattern. I have seen a small amount of "stress" in those answers, but nothing comparable to a stress pattern when the suspect lies on a relevant question.

As far as recorded material being analyzed by the Lantern, I personally am not a big proponent. I have had some success in analyzing audiotapes, but I find the charts much more difficult to analyze. I have used both cassette and DAT and I really don't see much of a difference between the two. They are both just as difficult for me to interpret. The patterns seem to appear much different that when a "live" test is administered. I do not feel that I can say that the taped material gives inaccurate readings, it may be just a personal preference on my part.

Another area of concern that I have concerning "live tests" is the possibility of interference. I have noticed that if I am conducting a test in a room, which contains a computer, there are noticeable differences in the patterns produced. I have shut the computer off and then asked the same question and received the same answer from the suspect, but the pattern is now different. Assuming that there was no other ambient noise during both times the question was asked, the patterns should be the same, except of course, if the interference was coming from the computer. On the same note, I also have noticed that possibly some interference caused from fluorescent lights. This should be an area of concern and perhaps more testing should be done to determine if the Lantern operates effectively under the above listed conditions.

One final and perhaps most important point I would make regarding the Lantern is the fact that you should not rely solely on the charts to make a determination if someone is "lying." I am not saying that Diogenes professes that this is a "lie detector", actually they profess the opposite. I am just saying that this should never be looked at as a "lie detector." I have truly found that it CANNOT detect lies. As you know, it DOES detect stress. Stress, however, does not always equate to a "lie." I have found in several cases that a person "fails", if you will, on all relevant/crime questions, but has been found to have not committed the crime.

I will close by saying that my experiences with the Lantern have been very positive, however, it cannot be looked upon as a "magic bullet." It is simply an investigative tool. Interrogation and the manner in which questions are formulated are very important. I truly believe that a person that is not strong in the interrogation area will not be as successful with it, as the person that possesses strong interrogation skills. There is much open to interpretation on the charts as far as I can see. It is very situational and again, can NEVER be determined a "lie detector."

James F. Masucci
Rome P.D. 05/17/00

These two reports reinforce the results of the technical testing, in that these systems do indicate stress. Caution should be taken when using these systems. They should only be used as investigator tool, and not total rely on these systems for a case conclusion.

8.0 CONCLUSION

After reviewing the three technical tests performed, it could be stated that these two VSA units do recognize stress. Although these systems state they detect deception, this was not proven. This study does show, from a number of speech under stress studies, that linear and non-linear features are useful for stress classification. Due to the lack of deceptive stress data available, classification of deceptive stress versus emotional stress or physical stress was not tested. This is a vital role in the detection and classification of stress. Many suspects are under an extreme amount of stress when being interrogated. Do these VSA systems actually differentiate between the different types of stress? This still needs to be proven.

It was shown, under test 1, that the Diogenes Lantern system detects stress via the amount of energy in the speech envelope. Even though this system performed well under the technical and field tests, it seems from an engineering point of view, that one feature, such as duration, glottal source factors, pitch distribution, spectral structure, or intensity, is insufficient to detect and classify deceptive stress. In the study under Dr. Hansen, it was shown that fusion of features help to increase the accuracy of stress classification.

It was proven that the systems tested will and do give the same response when the audio is recorded as opposed to live. The only criterion is when recording using a cassette player, set the AGC, this will prevent any audio clipping. To eliminate the possibility of this error, recording with a DAT is the safest way to go.

9.0 SUGGESTED FOLLOW ON

As it was stated, the biggest challenge that was encountered during this project, was the unavailability of sufficient deceptive stress data with ground truth. To make an accurate assessment of these systems, in respect to detecting deception, this data is needed. To develop this database three parties need to be involved. Walter Reed Army Institute of Research (WRAIR) will be tasked to collect/analyze a robust stress database, while evaluating deceptive stress vs. physiological and biochemical stress. The Department of Defense Polygraph Institute (DODPI) along with AFRL/Rome Research Site (RRS) and NLECTC/NE (Law Enforcement Analysis Facility (LEAF)) will collect deceptive stress data and test VSA systems simultaneously with polygraph under neutral/"crime subject" conditions. Personnel from the LEAF will be used because of their extensive VSA background.

For years, the Department of Justice and Law Enforcement Agencies in the United States have had only the polygraph technology as a "deception" indicator. With this recommend program, another "deception" indicator will be evaluated separately, and in a complimentary role with polygraph technology. AFRL/RRS will investigate this complimentary role, but in addition, possibly lay the groundwork for the future "fusion" of the two technologies, in an attempt to raise the confidence levels to the more acceptable standards of our justice system.

REFERENCES

[1] Allan Bell Enterprises, "Comparisons of Existing Lie-Detection Equipment," Unpublished.

[2] Allan D. Bell Jr., "The PSE: A Decade of Controversy," Security Management, March 1981, pgs. 63-73.

[3] " TVSA3 : Voice Stress Analysis Freeware" http://www.whatreallyhappened.com/RANCHO/POLITICS/VSA/truthvsa.html, August 1999.

[4] "The CCS Group" http://www.spyzone.com/catalog/index.html

[5] "National Institute for Truth Verification" http://www.cvsa1.com

[6] "Diogenes Company" http://www.diogenesgroup.com/

[7] "Trustech LTD." http://www.truster.com/

[8] Joseph F. Kubis, "Comparison of Voice Analysis and Polygraph as Lie Detection Procedures," U.S. Army Land Warfare Laboratory, LWL-CR-03B70, August 1973.

[9] Olof Lippold, "Physiological Tremor," Scientific American, Volume 224, Number 3, March 1971.

[10] D. H. VanDercar, J. Greaner, N.S. Hibler, C.D. Spielberger, and S. Bloch, "A Description and Analysis of the Operation and Validity of the Psychological Stress Evaluator," Journal of Forensic Sciences, January 1980, page 174-188.

[11] "Voice Stress Analyzer Kit" Instructions, XANDI Electronics Model No. XVA250.

[12] Merck Manual
John H.L.Hansen, Guojun Zhou and Bryan L. Pellom, "Methods for Voice Stress Analysis and Classification," Robust Speech Processing Laboratory, University of Colorado, July 1999.

[13] John J. Palmatier, "A Field Study to Test the Validity and Comparative Accuracy of Voice Stress Analysis as Measured by the CVSA: In a Psychophysiological Context," Research Proposal, The Michigan Department of State Police, Forensic Science Division, January 1997.

[14] Validation Program for Lie Detection Techniques Using Voice Analysis," U.S. Army Land Warfare Laboratory Purchase Order #DAAD05-69-M-5025, August 1969.

Appendix A

SPT – A numeric value describing the relatively high frequency range. Vericator associates this value with emotional stress level.

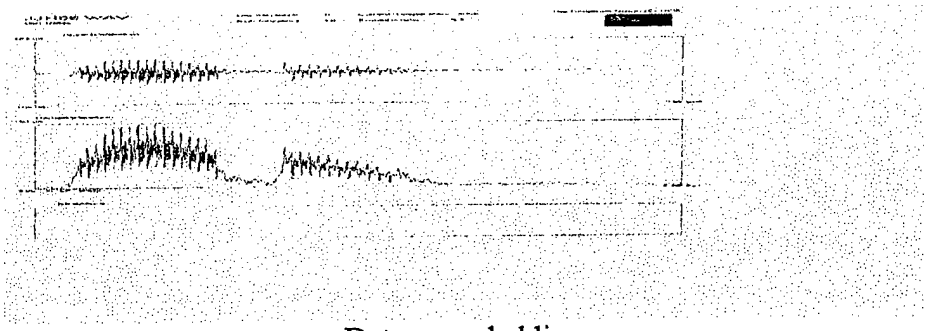
SPJ - A numeric value describing the relatively low frequency range. Vericator associates this value with cognitive stress level.

JQ - A numeric value describing the distribution uniformity of the relativity low frequency range. Vericator associates this value with global stress level.

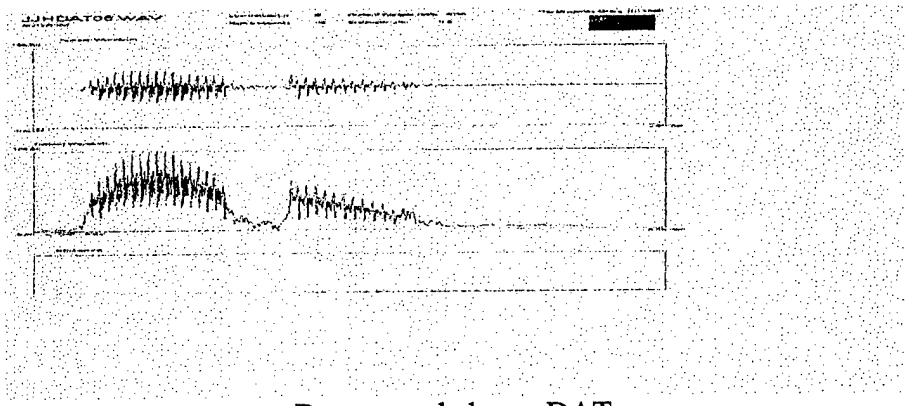
AVJ - A numeric value describing the average range of the relativity low frequency range. Vericator associates this value with thinking level.

SOS (SFLC) – **S**ay **O**r **S**top, a numeric value describing the changes in the SPT and SPJ values within a single sample sequence. Vericator associates this value with fear and the "breaking point" of the subject.

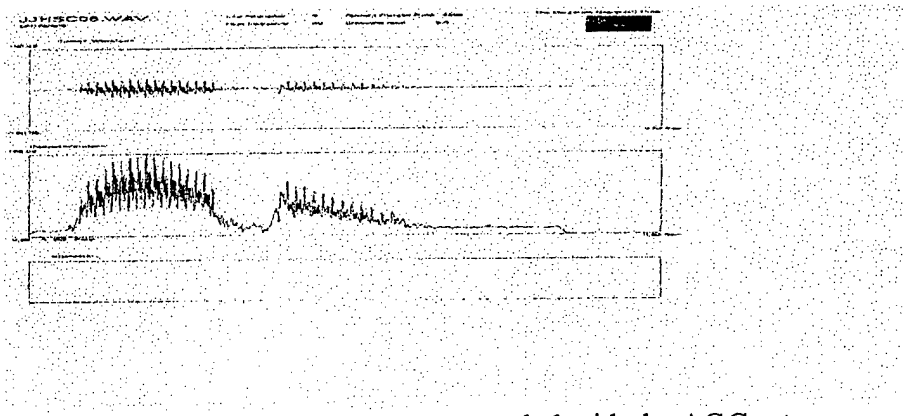
Appendix B



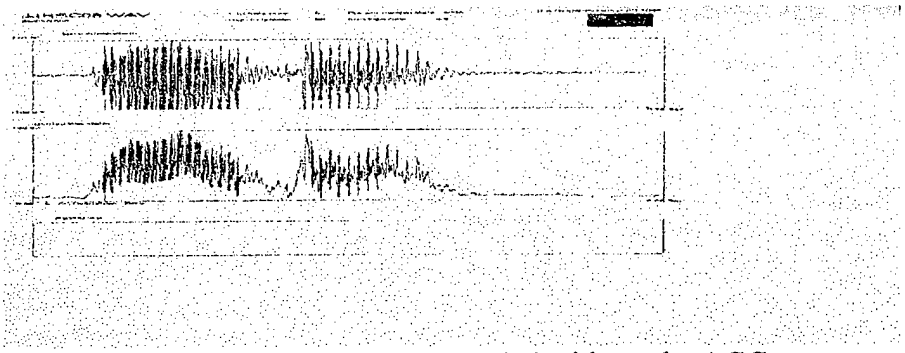
Data recorded live



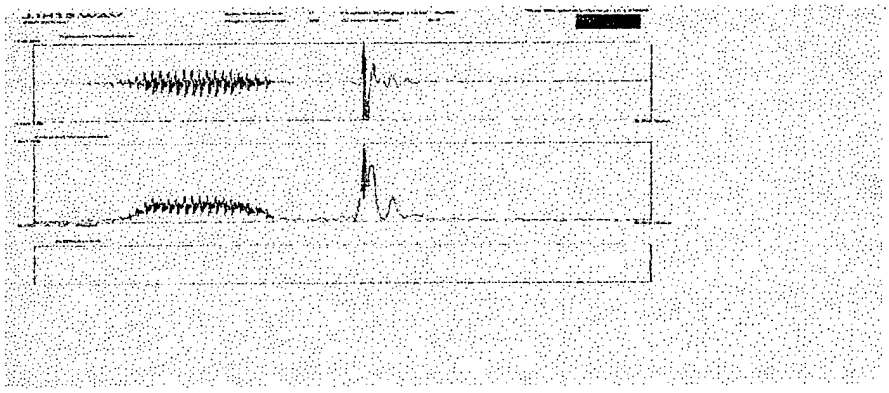
Data recorded on a DAT



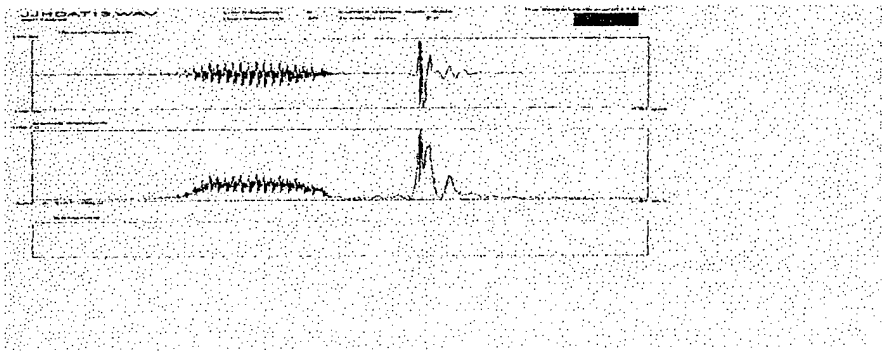
Data recorded on cassette recorded with the AGC set



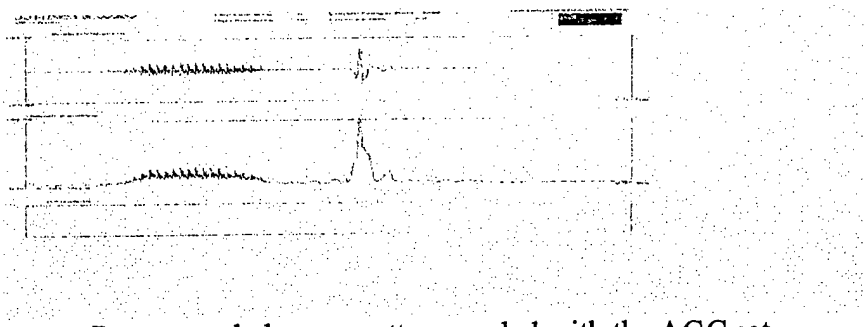
Data recorded on cassette recorded without the AGC set



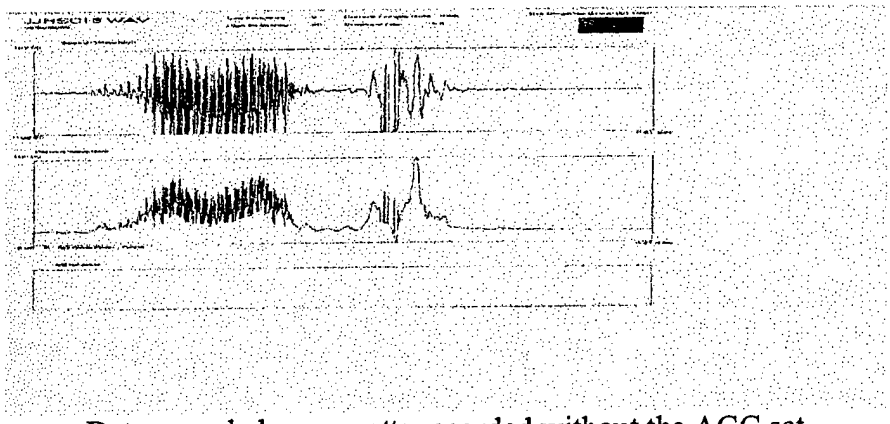
Data recorded live



Data recorded on a DAT



Data recorded on cassette recorded with the AGC set



Data recorded on cassette recorded without the AGC set

Appendix C

ANALYTICAL SYSTEMS ENGINEERING CORPORATION
(NOW ACS DEFENSE, INC.)
&
U.S. AIR FORCE RESEARCH LABORATORY
ROME, NY

METHODS FOR VOICE STRESS ANALYSIS AND CLASSIFICATION

John H.L. Hansen
Principal Investigator

Guojun Zhou
Bryan L. Pellom

Final Technical Report

RSPL-99-August

Project Period: August 1998 — July 1999

RSPL: ROBUST SPEECH PROCESSING LABORATORY
CENTER FOR SPOKEN LANGUAGE UNDERSTANDING
University of Colorado, Campus Box 594
Boulder, Colorado U.S.A.

Phone: (303) 735-5148 INTERNET: jhlh@cslu.colorado.edu
FAX: (303) 735-5072 http://cslu.colorado.edu/rspl/

1 Final Project Report

1.1 Executive Summary:

Current speech processing algorithms for classification and assessment of speaker stress which are designed to address DOD and law enforcement applications in such areas as automatic speech recognition, speaker identification, or gisting techniques for message sorting and translation lack the necessary signal processing capability to achieve reliable performance in emotional or task induced stressed environments. Unfortunately, in most applications requiring a man-machine interface, speaker monitoring, or analysis of subject interviews or telephone callers, it is specifically these high stress, emotional, deceitful, or emergency situations where reliable performance is critical. There have been much activity recently in the commercialization of voice stress analyzers for law enforcement applications. This study does not seek to directly prove or disprove these commercial systems, since in most cases the underlying details of their algorithms are typically revealed. Instead, we focus on features which have been used for stress assessment in both the linear and nonlinear speech processing domains. Those linear based speech features include: pitch, periodicity, jitter, glottal spectral slope, duration, intensity, formant locations, spectral structure as represented by the Mel-frequency cepstral parameters (MFCC), and the CVSA based measure. A number of nonlinear based features were also considered based on signal processing methods using the Teager Energy Operator (TEO). Our focus in this study was to use the measure based on an auditory critical-band frequency partition with temporal consistency represented by the autocorrelation envelope response across critical bands (i.e., the TEO-CB-Auto-Env feature). These features were implemented and evaluated using speech data from a number of corpora (i.e., SUSAS, SUSC-0, and ASEC-Stress data).

VOICE STRESS ANALYSIS:

During the project period, we completed three manuscripts, which summarize research conducted over the past two years on voice stress analysis for NATO IST/TG-1. The studies focused on analysis of fundamental frequency (pitch), duration, intensity, glottial spectral structure, and vocal tract formant characteristics. An analysis of vocal tract articulatory profiles under stress was also considered.

STRESS CLASSIFICATION:

In this area, we formulated an optimum detection algorithm for stress classification based on Bayesian Hypothesis Testing. Five speech production feature areas, originally investigated for analysis of speech under stress, were evaluated for optimum stress detection using the SUSAS speech under simulated and actual stress database. Results showed that pitch (fundamental frequency) was the best feature for stress classification (equal error rate EER=11.4%), followed by individual phone class intensity and duration (ERR=23.1% and ERR=30.8%), and to a lessor degree glottal spectral slope (EER=32.2%). Individual formant locations (first and second) were not reliable features for stress classification (ERR=45.5% and 46.3%).

Next, we considered features derived from signal processing scheme based on the nonlinear Teager Energy operator (TEO). This operator assumes that a speech resonance to be modeled by an AM/FM component. While several TEO based stress classification measures were proposed, here we present the TEO-CB-Auto-Env. This measure is based on an auditory based critical band frequency partition, followed by an autocorrelation envelope estimation. The feature represents the time-domain correlation of AM/FM based structure across a partitioned frequency band. Results presented for stress detection show that the TEO-CB-Auto-Env measure (mean classification rate of 94.2%) outperforms pitch (mean classification rate of 88.5%), vocal tract spectral characteristics as represented by the mel-frequency cepstral coefficients (MFCCs) (mean classification rate of 89.6%

STRESS ASSESSMENT:

During the project period, we extended the application of stress classification based measures to the problem of stress assessment. To determine the usefulness of the nonlinear TEO-CB-Auto-Env measure for stress assessment, we considered an evaluation of SUSC-0 speech corpus from NATO IST/TG-01. This evaluation focused on the *Mayday2* domain, which involved voice communication between an aircraft pilot and controller during an emergency where is engine fails. The TEO based measure was shown to follow the perceived level of stress in the extracted voice recordings based on a secondary informal listener evaluation. This result is meaningful, since the anchor neutral/stress models used in the assessment were based on speech data from the SUSAS stress database (i.e., open speakers and open training speech material).

COMMERCIAL/CONVENTIONAL 'VSA' FEATURES:

A number of commercial voice stress analyzers have recently appeared on the market. These methods are based on some form of speech signal processing to extract excitation information related to small microtremors which are believed to be associated with the laryngeal muscles during voiced excitation. Physiological tremor is produced through repetitive movement of muscle contraction and relaxation. Slow tremor occurs at rates between 3-5Hz, while rapid tremor occur between 6-12Hz. Benign hereditary tremor is a fine-to-coarse slow tremor that usually effects the hands, head, and voice. Such tremor generally increases with age, and in some cases (some families), ingestion of small amounts of alcohol markedly suppresses the tremor. Other forms of tremor in voice are associated with neurological speaker changes such as the resting tremor seen in Parkinson's speech. There are many causes of tremor, which include medical illness, drugs, stress, and brain disorders such as multiple sclerosis.

The focus here is on how stress impacts the laryngeal muscles during normal production of speech, and whether speech processing algorithms/systems are able to extract and quantify such information if it exists. In our study, we focused on excitation features which include (i) normalized pitch, (ii) periodicity, and (iii) jitter. These features have long been used in the medical field for assessing changes in speech production due to pathology such as vocal fold cancer, vocal fold nodules, or other physically based change in the structure or movement of the vocal folds during phonation. Here, we evaluated these features for the purpose of voice stress classification using a Gaussian mixture model (GMM) classifier. In the evaluations, we considered a range of GMM classifier mixture weights, training iterations, static features with and without first and second order derivative features, and combinations with spectral parameters. The best GMM classifier included all three excitation features with first and second order derivatives, a feature trained variance threshold of 0.001, 64 Gaussian mixtures, and at least some form of overall vocal tract spectral structure if the data is available. It should be noted that these methods are effective only if the speaker conveys changes in his excitation, or in the laryngeal muscles associated with microtremors. A number of useful studies have considered the use of computer voice stress analyzers for the purpose of detection of deception. Some of these include:

D. VanDercar, J. Greaner, N. Hibler, C. Spielberger, S. Bloch, "A description and analysis of the Operation and Validity of the Psychological Stress Evaluator," *Journal of Forensic Sciences*, vol. 25, no. 1, pg. 174-188, Jan. 1980.

F. Horvath, "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, pp. 338-344, 1978.

O. Lippold, "Physiological Tremor," *Scientific American*, vol. 224, no. 3, pp. 65-73, 1971.

In addition to these references, there are a number of commercial voice stress analysis systems (e.g., Israeli system called *TRUSTER*, Psychological Stress Evaluator (PSE) by Verimetrics, Computerized Voice Stress Analyzer (CVSA), and others). Our findings suggest that if the input speaker does in fact produce microtremors in their laryngeal muscles during voiced speech production, then the simple filtering operation proposed in PSE (Bell, et. al, 1976), CVSA, and others, can extract the presence or absence of such tremor. It has been suggested that if the natural tremor is absent, then the speaker is experiencing stress, and if the fluctuations are present the person is not experiencing stress. Again, extreme caution should be exercised in using these devices because it is not necessarily true that the muscle tremor associated with stress or deception will *always* effect those laryngeal muscles using during phonation in the same manner for all speakers. This is a well known and documented observation in the area of vocal fold cancer detection (Hansen, Gavidia-Ceballos, Kaiser, 1998), since it is possible that a physiological change in the vocal folds (a cancer growth, or muscle change/paralysis) may not always impact the normal mucosal wave, and therefore will not be represented in the sound pressure wave which excites the vocal tract. We discuss a number of these issues in the summary and conclusions section.

STRESS ASSESSMENT: Mount Carmel Law Enforcement Evaluation

In the final section of this study, we consider the application of the voice excitation features used by commercial voice stress analyzers for stress assessment of the 911 audio recordings obtained from the Law Enforcement encounter with an extremist sect in Mount Carmel. These recordings were between the sect leader and a 911 operator during the FBI encounter. For analysis, we considered normalized pitch, periodicity, jitter, and a software implemented version of the commercial CVSA system. The resulting feature profiles were compared with feature profiles

obtained from an evaluation of speech data from the SUSAS speech unders stress database. The profiles for normalized pitch and periodicity did not appear to be reliable indicators of speaker stress. The CVSA profile did show some of the structure which were expected for Mt. Carmel data, but was not as consistent for SUSAS actual stressed speech (this could be explained because that speech portion of SUSAS was collected during amusement park roller coaster rides which could have introduced physical vibrations during speech production). An extensive evaluation of the entire series of sentences for stress assessment using pitch, spectral MFCC features, and the TEO-CB-Auto-Env measure showed that pitch and the new TEO-CB-Auto-Env measure produced more consistent assessment scores. A number of issues regarding successful stress classification and assessment using either traditional excitation features or nonlinear speech features must be addressed to achieve successful voice stress analysis performance. Ultimately, the success of the measure rests on how the speaker imparts, either consciously or subconsciously, stress in their speech production process (either through controlled airflow from the lungs, muscle control of the vocal system articulators, choice of vocabulary). It is suggested that more success could be achieved if the subjective impression of the operator could be reduced for commercial VSA devices. Further training data for model adaptation, and establishing well recognized anchor neutral/stress models for a given speaker in context, should ultimately produce more reliable performance. At best, the available commercial systems should be used with caution if they are to be applied.

1.2 Outline of Report

The outline of this report is as follows. In Section 2, we provide a review of the literature in speech under stress. Next, in Section 3 we present a brief overview of results from analysis of speech production under stress which include pitch, speech duration, speech intensity, glottal spectral structure, and formant structure. Section 4 discusses methods considered for stress classification. This section focuses on previous approaches, Bayesian stress classification, linear feature classification, nonlinear based features. Section 5 considers the use of stressed classification features for stress assessment using actual stressed speech from a pilot emergency (MAYDAY2 portion of the SUSC-0 corpus from NATO). This worked focused on normalized

pitch, spectral structure using MFCCs, and the nonlinear TEO-CB-Auto-Env measure. Finally, Section 7 presents a series of probe evaluations of speech data from Mount Carmel. This represents speech data from a high stress law enforcement encounter. Since the available speech was limited, the analysis was restricted to comparisons of feature profiles for linear features, and assessment evaluations using anchor models trained with SUSAS stressed speech data. In the Appendix (Section 9), we identify the software, which has been implemented and will be delivered to the sponsor as part of this project.

2 Speech Under Stress: Review of the Literature

Stress is a psychological state that is a response to a perceived threat or task demand and is accompanied by specific emotions (e.g., fear, anxiety, anger). Initial investigations of verbal indicators of stress have focused on identifying speech markers of stress (e.g., stuttering, repetition, tongue-slip). Psychiatrists agree that verbal markers of stress range from highly visible to invisible markers as perceived by the listener (Goldberger and Breznitz, 1982), and that these markers are continuously monitored both consciously and subconsciously by the speaker and thus are prone to correction.

2.1 Acoustic Correlates of Stress and Emotion in Speech

A number of studies have considered analysis of speech under simulated and actual stress conditions (see Table 1), though changes in speech characteristics remain unclear. Thus far, most research has been limited in scope, often using only one or two subjects and analyzing a single parameter (often f_0). It is not unusual for researchers to report conflicting results, due to differences in experimental design, level of actual or simulated stress, or interpretation of results. For example, some studies concentrate on analysis of recordings from actual stressful situations (Kuroda, et al. 1976; Simonov and Frolov, 1977; Streeter, et al., 1983; Williams and Stevens, 1972). There is usually little doubt as to the presence of stress in these recordings, however a quantitative analysis can only be carried out if recordings of the talker speaking the same utterances under stress-free conditions is available. In addition, some researchers argue that speakers in these situations may experience several emotions simultaneously, (e.g., the Hindenburg announcer most likely experienced combinations of fear, grief, and anxiety). Another group of studies have been performed using simulated stress or emotions (Hecker, et al., 1968; Hollien and Hicks, 1981a, 1981b; Williams and Stevens, 1972). This offers the advantage of a controlled environment, where a single emotion can be examined with little background noise. In some cases, variable task levels of stress have been used. Other advantages include larger data sets with multiple speakers. The major disadvantage in these studies have been the reduction in task stress levels. In addition, studies using actors may produce exaggerated caricatures of emotions in speech.

In previous work, Williams and Stevens, (1972), and Hecker, et al.(1968) found that f_0[1] to be the acoustic property most sensitive to the presence of stress. There are several reasons why changes in f_0 with time provide information on emotional state. For example, respiration is frequently a sensitive indicator in certain emotional situations. When an individual experiences a stressful situation, his respiration rate increases. This presumably will increase subglottal pressure during speech, which is known to increase f_0 during voiced sections (Pickett, 1980). An increased respiration rate also leads to shorter durations of speech between breaths, which would affect the temporal pattern (articulation rate). The dryness of the mouth found during situations of excitement, fear, anger, etc. can also effect speech production (e.g., muscle activity of larynx and condition of vocal cords). Muscle activity of the larynx and vibrating vocal cords directly affect the volume velocity through the glottis, which in turn affects f_0. Other muscles, (for example those controlling tongue, lips, jaw, etc.) shape the resonant cavities for sound and therefore do not have a direct influence on f_0.

2.2 Analysis Using Simulated Stress or Emotion

Here, analysis of studies using simulated stress or emotion are considered first (see Table 1). Here, we place emphasis on the study by Williams and Stevens (1972) since they considered analysis of recordings from both simulated[2] and actual[3] emotional environments. Hicks and Hollien (1981a,b) simulated stress by using mild electrical shock. Hecker et al., (1968) simulated stress by having subjects perform a timed arithmetic task.

Fundamental frequency f_0 contours and f_0 variability were analyzed for anger, sorrow, and fear by Williams and Stevens (1972). For fear, the f_0 contour departed greatly from neutral, while for anger the contour was generally higher throughout with one or two syllables characterized by large peaks. Hicks and Hollien (1981a,b) found similar increases in f_o.

[1] Most early studies on speech under stress consider fundamental frequency, but use the term pitch. Since pitch is a perceptual quantity, our research here will focus on fundamental frequency.
[2] Simulated recordings consisted of data from actors simulating fear, anger, sorrow, and neutral emotions.
[3] Actual recordings consisted of data from the radio announcer during the Hindenburg disaster.

However, Hecker et al. (1968) observed conflicting results (some subjects increased, while others decreased f_0).

Mean articulation rate in syllables/second was determined for the three emotions considered by Williams and Stevens (1972). Results from fastest to slowest were neutral, anger, fear, and sorrow. Hicks and Hollien (1981a,b) also observed similar results.

Speech intensity or vocal effort per unit time, during voiced sections was considered by Hicks and Hollien (1981a,b) and Hecker et al. (1968), although inconsistent results occurred across test phrases. Pisoni et al. (1985), Summers et al. (1988) investigated acoustic-phonetic correlates of speech produced in noise (also called the Lombard effect (Lombard, 1911)). With subjects speaking in quiet and 90 dB SPL white masking noise, results showed an increase in overall amplitude of vocalic sections, increased duration, increased average f_0, and reduced spectral tilt. Junqua (1993, 96) also performed analysis on Lombard effect speech and concluded that female speakers seem to be more intelligible than male speakers. Rostolland (1982a,b) performed acoustic and phonetic studies for shouted speech and observed reduced intelligibility with a raised f_0 contour.

In other investigations, Lieberman and Michaels (1962) had subjects simulate eight emotional states. Their approach was to select a parameter as an emotion relayer, extract that parameter, and observe whether the resulting sound could correctly be identified as the simulated emotion by listener groups. While only characteristics of pitch and amplitude were considered, results showed that fear was highly identified using only amplitude information with constant pitch.

2.3 Analysis Using Actual Stress or Emotion Situations

A comparison of results from actual stressful recordings is somewhat difficult, due to varying parameters measured and levels of stress experienced. However, considering such studies are important, since the analysis may help verify experimental procedures and results from simulated studies.

Kuroda et al. (1976) analyzed tape recordings of pilots with varying mission experience in actual aircraft accidents. Their analysis consisted of finding a parameter related to pitch, termed the vibration space shift rate (VSSR) from speech spectrograms. Their ultimate conclusions showed as stress increased so did f_0. A more recent study by Ruiz, et al. (1996) considered time and frequency-domain analysis of emergency aircraft cockpit recordings.

Simonov and Frolov (1977) analyzed communications of cosmonauts at various flight stages. Analysis consisted of monitoring heart rate and the spectral centroid of the first vocal tract formant. Though general trends were noted, their summary emphasized the need of further research.

Streeter et al. (1983) carried out a more complete analysis of a telephone conversation between a system operator (SO) and his superior chief (CSO) prior to the 1977 New York City blackout. Analysis consisted of pitch, amplitude, and timing measurements. An attractive feature of the data was the increased situational stress throughout the hour-long conversation. Results were somewhat conflicting since it appeared SO was passing decision making authority to CSO during the emergency. Results showed that listeners referred to a vocal stress stereotype, which includes: elevated pitch and amplitude, and increased variance in these vocal cues.

Finally, Williams and Stevens (1972) performed analysis of the recorded radio announcer during the Hindenburg disaster. In an effort to justify results from their simulated emotions, they had actors recreate the announcer's message. Results were not entirely consistent, though increased average f_0 along with tremor were observed for both, with larger variations for the actor. This would indicate that the actor's emotions to a certain extent were overemphasized. This, as well as other previous studies on speech under stress are summarized in Table 1.

2.4 Voice Stress Analysis for Law Enforcement

Most of the studies discussed in Section 2.2 and 2.3, and summarized in Table 1, deal with speech produced in either emotional or task induced stressful environments. There is another area of analysis of speech under stress, which deals with voice tremor and how it applies to detection of deception for law enforcement applications. The study by Cestaro (1995) was an

extensive evaluation of the commercial Computer Voice Stress Analyzer (CVSA). In that study, two experiments were designed to validate the underlying theory of CVSA, and second to examine the accuracy of CVSA with traditional polygraph instrument for the problem of stress detection due to deception. He simulated the CVSA signals electronically using signal generators in order to have careful control over what commercial have been made for CVSA. His findings show that CVSA was less successful and accurate than a polygraph (38% versus 62%). His results suggest that there may be a systematic and predictable relationship between voice patterns and the stress related to deception.

Another study by VanDercar, et. al, (1980) detailed an evaluation of the psychological stress evaluator (PSE) as a commercial system for representing different levels of speaker stress. They measured PSE profiles in addition to heart rate and State Trait Anxiety Inventory (STAI) scores during relaxed and high stress (through the threat of electric shock). When the potential for stress was high, PSE, STAI, and heart rate measures all reflected different levels of stress and were significantly correlated with each other. A second study with reduced stress levels failed to show the reliability of PSE. It was suggested that the lower levels of stress were a factor in the difference in performance for the second experiment. In later sections, we consider a computer implementation of the stress classification feature within the CVSA instrument. Evaluations are performed for SUSAS and Mt. Carmel stressed speech recordings (this data will be discussed later).

Table 1: A summary of studies on speech under simulated and actual stress conditions.

Summary of Speech Under Stress Studies		
Simulated Stress	**Speech Analysis Areas**	**Stress/Emotion**
Lieberman & Michaels (1962)	Pitch & Amplitude (Listener Assessment)	Simulated Stress: Simulated Emotion
Hecker, Stevens, et al (1968)	Mean Pitch Speech Level Spectogram Comparison	Simulated Stress: Timed Arithmetic Task
Hicks & Hollien (1981)	Mean Pitch Mean Intensity Speech Rate	Simulated Stress: Mild Electric Shock
Rostolland (1982)	Acoustic Analysis	Simulated Stress: Shouted Speech
Pisoni, et al. (1985)	Pitch Duration General Spectral	Simulated Stress: Lombard Effect
Stanton, et al (1988)	Pitch Duration Frequency Characteristics	Simulated Stress: Loud and Lombard Effect
Junqua (1993)	Pitch Duration Frequency Analysis	Simulated Stress: Lombard Effect
Actual Stress	**Speech Analysis Areas**	**Stress/Emotion**
Kuroda, et al. (1976)	Pitch (VSSR)	Actual Stress: 14 Pilot Emergency Cockpit Recordings (8 Fatal)
Simonov & Frolov (1977)	First Formant Analysis Heart Rate	Actual Stress: Cosmonaut Flight Recording Analysis
Streeter, et al. (1983)	Pitch Speech Level Timing Measurements	Actual Stress: Telephone Call Analysis of Con.Ed. New York City Backout (1977)
Simulated & Actual	**Speech Analysis Areas**	**Stress/Emotion**
Williams & Stevens (1972)	Pitch Contours Pitch Variability Spectrogram Comparison Avg. Spectral Content Mean Articulation Rate	Simulated Stress: Method Actors: Fear, Anger, Sorrow Simulated Hindenburg Announcer Actual Stress: Hindenburg Announcer
Hansen (1988)	Pitch Glottal Source Duration Intensity Vocal Tract Spectrum (+200 Speech Features)	Simulated Stress: Fast, Slow, Loud, Soft, Clear, Angry, Question, Lombard Effect, Computer Response Task Stress Actual Stress: Roller-Coaster Ride Speech, Psychiatric Emotional Analysis

3 Acoustic/Phonetic Analysis of Speech under Stress

In order to perform an in depth study, a comprehensive speech under stress database, entitled SUSAS (Speech Under Simulated and Actual Stress) was formulated (Hansen, 1988; Hansen and Bou-Ghazale, 1997). The database is partitioned into five domains, encompassing a wide variety of stresses and emotions. A total of 32 speakers (13 female, 19 male, with ages ranging from 22 to 76 were employed to generate in excess of 16,000 utterances. Table 2 illustrates the various domains present in the database. The vocabulary consists of 35 aircraft communication words containing a number of subsets that are difficult for recognition systems.

Table 2: The SUSAS Speech under Simulated and Actual Stress Database.

Susas Database Speech Under Simulated and Actual Stress				
Domain	**Type of Stress or Emotion**	**Speakers**	**County**	**Vocabulary**
	Simulated Stress Slow Soft Fast Loud Angry Clear Question	9 Speakers (All Male)	8820	35 Aircraft Communication Words
Single Tracking Task	Calibrated Workload Tracking Task: Moderate & High Stress Lombard Effect	9 Speakers (All Male)	1890	35 Aircraft Communication Words
Dual Tracking Task	Acquisition & Compensatory Tracking Task: Moderate & High Stress	8 Speakers (4 Male) (4 Female)	4320	35 Aircraft Communication Words
Actual Speech Under Stress	Amusement Park Roller-Coaster Helicopeter Cockpit Recordings (G-Force, Lombard Effect, Noise, Fear, Anxiety)	9 Speakers (4 male, 3 Female) (2 Male)	500	35 Aircraft Communication Words
Psychiatric Analysis	Patient Interfiews: (Depression, Fear, Anxiety, Angry)	8 Speakers (6 Female) (2 Male)	600	Conversational Speech:Phrases & Sentences

3.1 Analysis Overview of Speech Under Stress

In this section, we summarize the analysis conducted on speech from (i) *Simulated stress or emotion*, and *Simulated workload task or Lombard effect*}, followed by (ii) *probe studies using Actual speech under stress*. In this context, *simulated* conditions refer to areas where talkers were asked to either speak in a prescribed emotional manner, or perform some computer response task while uttering speech. For these domains, control of experimental conditions and

environmental factors were possible (e.g., vocabulary, task difficulty). *Actual* conditions refer to areas where talkers are in environments similar to real stressful situations. These domains differ from simulated conditions in their lack of control in experimental and environmental factors (e.g., varying noise levels, task difficulty, vocabulary choice).

To our knowledge, there is no singly reliable acoustic indicator of psychological stress. There has been a lack of consistent results in past research efforts. After considering experimental design and analysis, it is apparent that past approaches to stress analysis suffer from one to five of problems summarized as follows; (i) analysis based on too few speakers, (ii) analysis based on too few utterances, (iii) analysis based on a limited set of parameters with no consideration within speech sound classes, (iv) no statistical analysis to determine if changes are statistically significant, (v) no confirmation of simulated results with those from actual recordings. Here, we address these problems in the context of fundamental frequency, duration, and intensity, glottal source, and spectral factors. The analysis included extensive parametric and non-parametric statistical tests (see Hansen, 1998a, 1998b).

3.2 Analysis of Fundamental Frequency

The first area considered for stress evaluation involves characteristics of fundamental frequency f_0, including contours, mean, variability, and distribution.

A subjective evaluation of more than 400 f_0 contours was conducted across all stress conditions (sample contours[4] are shown in Fig. 1). The overall shape of the contours for fast and slow speech did not change appreciably. Angry and loud contours had much higher variability than neutral, with angry the highest mean and variability of all stress conditions considered. f_0 contours of soft speech were almost always smoother than neutral. Speech under Lombard effect had a slight elevated mean, but the contours appeared similar in shape. Contours for moderate and high workload task conditions were similar to neutral.

[4] Here, the f_o contours for the word "histogram" are shown continuous, because the timing of the unvoiced obstruent /s/ varies during production across the stress styles, so endpoints of the contours are joined in this example, and all statistical analysis was performed directly on frame based fundamental frequency values.

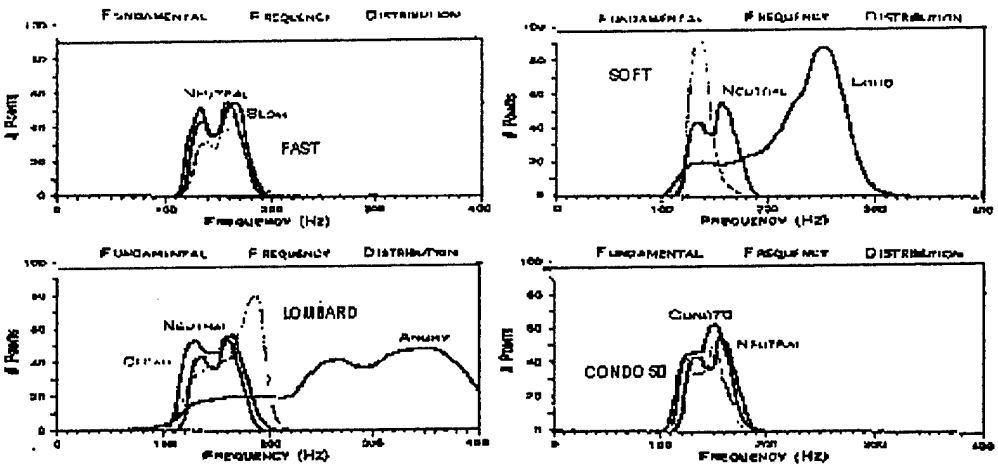


Figure 1: Distribution of fundamental frequency based on fundamental frequency estimates from neutral, angry, fast, slow, loud, soft, clear, Lombard effect, moderate (50%), and high (70%(workload condition speaking styles.

Next, we consider differences in mean, variance, and distribution of pitch (f_0). Since a number of statistical tests performed assume sample variables to be Gaussian distributed, a comparison of f_0 distribution contours was performed (see Fig.1). We are primarily interested in seeing whether the distribution shape differs substantially from Gaussian. f_0 distributions for neutral, clear, slow and fast have similar shapes with a bimodal concentration. Negative kurtosis values (from Table) confirm distributions which are more flat as compared to Gaussian. Lombard and loud f_0 distributions had similar shape, along with values of skewness and kurtosis, though the range of loud f_0 was wider. The soft f_0 distribution was highly concentrated with a very small variance, which was confirmed by a large positive kurtosis value suggesting a peaked distribution. Of the f_0 distributions considered, loud and angry styles were judged not to be Gaussian. Angry resulted in a very irregular shaped distribution, with large concentrations towards higher frequencies.

Finally, below we summarize some of the key findings for analysis of pitch mean and variance, based on speech data from the SUSAS database.

Summary of Key Results: Mean Fundamental Frequency vs. Stress

- The position of mean f_0 from highest to lowest versus speaking style is shown below.

Fundamental Freq.	Condition	Shift in mean f_0
Highest	Angry	+73% stat. Significant
	Loud	+48% stat. Significant
	Lombard	+10% stat. Significant
	Clear	+4% stat. Significant
	Fast	+6% stat. Significant
	Neutral	
	Slow	-2%
	Task Condition 70%	-2%
	Task Condition 50%	-3%
Lowest	Soft	-5% stat. significant

- Mean f_0 values may be used as significant indicators for speech in soft, fast, clear, Lombard, angry, or loud styles when compared to neutral conditions.
- Loud, angry, and Lombard mean f_0 are all significantly different from all other styles considered.
- Mean f_0 was not a significant indicator for moderate versus high task workload conditions.
- Speech under Lombard effect gave mean f_0 values most closely associated with f_0 from fast and clear conditions.
- Changes in mean f_0, based on Student's t-tests, appears to be a consistent and reliable stress indicator over a wide variety of conditions.

Summary of Key Results: Variance of Fundamental Frequency vs. Stress

- The position of f_0 variability from highest to lowest versus speaking style is shown below.

Pitch	Condition	Shift in Standard Deviation
Highest	Angry	+506% stat. significant
	Loud	+213% stat. significant
	Lombard	+55% stat. significant
	Clear	+59% stat. significant
	Slow	+19% stat. significant
	Fast	+2%
	Task Condition 50%	+3%
	Task Condition 70%	+2%
	Neutral	
Lowest	Soft	28% stat. significant

- Variance of f_0 values may be used as significant stress indicators for speech in soft, loud, angry, clear, or Lombard styles when compared to neutral conditions.
- Soft and loud f_0 variance are significantly different from all styles considered.
- Pitch variance was not significantly different for moderate versus high task workload conditions.
- Pitch variance was unreliable for slow and fast stress conditions.
- Pitch variance for clear and Lombard conditions are similar, but different from all other styles considered.

3.3 Analysis of Duration

In order to address duration in speech under stress adequately, analysis was partitioned into stress relayers across four areas. The first two focused on overall word and individual speech class (vowel, consonant, semivowel, and diphthong) durations. Third, analysis within speech classes provided detailed indicators of duration shifts between classes. Fourth, because overall word duration may supersede the requirements of lengthening consonantal periods (or semivowel, diphthong periods), several duration ratio measures are proposed.

Several comments concerning durational effects caused by prosodic features may help explain the durational variation caused by stress. There have been a multitude of studies investigating durational variations which arise from prosodic conditions (Fry, 1955; Creelman, 1962; House, 1962; Perkell and Klatt, 1986). A number of more recent studies have considered timing and height of pitch contours for female speakers (van Santen and Hirschberg, 1994), phone/syllable duration and timing representations for text-to-speech synthesis (van Santen, 1995), pitch and duration in signaling emotion (Ofuka, Campbell, et al., 1994), and segment duration in hidden Markov model speech recognition (Levinson, 1986; Wang, et al. 1996). Basic data and initial prosodic rules, which govern consonant and vowel duration, can be found in studies by Klatt (1973,76), Fry (1955), House (1962), and Umeda (1975,77). Duration patterns have also been studied in an attempt to arrive at principles of motor organization by Lindblom (1963), Lindblom, Lyberg, and Holmgren (1977), and Kohler (1986). Barnwell (1971) was the first to identify a limit to the temporal compressibility of vowels when they are followed by an unvoiced consonant and/or by an additional syllable. In a later study, Klatt (1973) expressed this incompressibility in a formula, and applied it as a rule to adjust consonant durations for various shortening effects (Klatt, 1976).

Below, we summarize results from the statistical evaluation of mean and variance of word and phoneme class duration.

Summary of Key Results: Mean Duration vs. Stress

- Mean duration from highest to lowest versus talking style for all speech classes (* = significant with respect to neutral)

Shift in Mean Duration

Word		Vowel		Consonant		Semivowel		Diphthong	
SL	+73%	SL*	+84%	C*	+84%	SL*	+112%	SL*	+94%
C*	+39%	A*	+69%	SL*	+52%	LM*	+63%	A*	+64%
A*	+38%	L*	+58%	SO	+24%	A	+42%	L*	+53%
L*	+36%	C	+26%	C7	+22%	C	+39%	LM	+30%
LM*	+20%	LM	+24%	C5	+12%	L	+27%	SO	+9%
SO	+7%	N		LM	+4%	C5	+20%	C	+4%
C_7	+5%	SO	-8%	L	+3%	SO	+19%	N	
C_5	+1%	C_5	-8%	N		C_7	+14%	C_7	-7%
N		C_7	-8%	A	-12%	N		C_5	-8%
F*	-26%	F*	-28%	F*	-27%	F	-27%	F	-27%

- Mean word duration values may be used as significant indicators for speech in slow, clear, angry, loud, Lombard, or fast styles when compared to neutral conditions
- Slow and fast mean word duration are all significantly different from all other styles considered
- Clear mean consonant duration was significantly different from all styles except slow
- Word and phoneme class duration are not significant indicators for moderate vs. high task workload conditions

Summary of Key Results: Duration Variance vs. Stress

- Duration variance from highest to lowest versus speaking (* = significant with respect to neutral)

Shift in Duration Variance

Word		Vowel		Consonant		Semivowel		Diphthong	
SL*	+173%	A*	+191%	C*	+456%	A*	+1045%	SL*	+324%
A*	+128%	SL*	+166%	SL*	+294%	SL*	+942%	A	+112%
C*	+122%	L*	+141%	L*	+106%	LM*	+531%	L	+70%
L	+56%	C*	+115%	A*	+83%	C*	+370%	LM	+6%
LM	+32%	LM	+65%	C7*	+78%	L*	+326%	C7	+3%
N		N		SO*	+63%	C5	+150%	C	+1%
SO	-11%	C5	-4%	LM	+44%	C7	+106%	N	
C5	-11	C7	-4%	C5	+39%	SO	+91%	C5	-27%
C7	-22%	SO	-22%	N		F	+45%	SO	-66%
F	-33%	F*	-54%	F*	-39%	N		F	-70%

- Duration variance increased for slow speech in all domains (word, vowel, consonant, semivowel, diphthong
- Duration variance decreased for most domains under fast stress condition
- Duration variance significantly increased for angry speech
- Duration variance generally increased for loud speech, but was mixed for soft speech
- Clear consonant duration variance was significantly different from all styles
- Duration variance is not a significant indicator for moderate versus high task workload conditions

Since overall word duration may supersede requirements for lengthening of consonants (or other speech classes), several duration ratio measures were proposed. The following three[5] ratios were considered; i) a consonant versus vowel duration ratio (CVDR), ii) a consonant versus semivowel duration ratio (CSVDR), and iii) a vowel versus semivowel duration ratio (VSVDR),

$$CVDR = \frac{d_{consonant}(neutral)}{d_{vowel}(neutral)} \tag{1}$$

$$CSVDR = \frac{d_{consonant}(neutral)}{d_{semivowel}(neutral)} \tag{2}$$

$$VSVDR = \frac{d_{vowel}(neutral)}{d_{semivowel}(neutral)} \tag{3}$$

where duration values $d_{class}(stress)$ are for a particular phoneme class and stress condition. It is suggested that such ratios can be used to determine directions in which speakers vary their duration patterns under stress. By using neutral ratios as baseline values for comparison, one can determine how phone class duration varies for individual stress styles.

[5] Duration ratios with respect to dipthongs were not considered, due to the limited number of examples available in the Talking Styles stress domain.

CVDR and CSVDR suggest that there is a shift in percentage time spent in vowels and semivowels towards consonants for soft, clear, and to a lesser degree the two task conditions. CVDR and VSVDR also revealed increased vowel duration at the expense of consonant and semivowel portions for angry and loud speech. It is difficult to get a clear picture of the global changes in duration from simply comparing the duration ratios. Therefore, a pictorial representation of global duration shifts is presented in Fig.2. A bar graph, proportional to average word length from the SUSAS database is shown for each stress condition. The percentage of vowel, semivowel, and consonant duration with respect to an overall average word duration is also shown within each shaded section. The percentage is simply the ratio of average phoneme class duration to that of an average word duration assuming one vowel, consonant, and semivowel. All calculations are based on tabulated values. As an example, the 24% consonant duration for neutral was obtained by assuming an ideal stressed word with one phoneme from each class as follows,

$$word = vowel + semivowel + consonant \tag{4}$$

$$\bar{d}_{word_{neutral}} = \bar{d}_{vowel}(neutral) + \bar{d}_{semivowel}(neutral) + \bar{d}_{consonant}(neutral) \tag{5}$$

$$295 = 165 + 59 + 71 \qquad \text{(in msec.).} \tag{6}$$

The consonant percentage is simply the ratio of the average consonant to ideal word duration.

$$Percent_{consonant} = \frac{\bar{d}_{consonant}(neutral)}{\bar{d}_{word}(neutral)} \tag{7}$$

$$24\% = \frac{71}{285} x100 \tag{8}$$

The arrows in Fig.2 indicate significant shifts in duration based on CVDR, CSVDR, and VSVDR. As an example, angry speech results in significant increases in vowel duration at the expense of semivowel and consonant duration. It is apparent from the results presented here the presence of stress influences overall and individual duration characteristics.

3.4 Analysis of Intensity

The control of vocal intensity is based on adjustments of laryngeal and subglottal variables. In addition, past research on the effects of intensity for speech intelligibility has also served to improve the knowledge of how speakers vary intensity in typical speech production. An analysis of consonant strength and precision was performed by House, et al.(1965). In this investigation, consonant-vowel amplitude ratios (CVAR) were measured for two speakers differing in intelligibility as measured by the Modified Rhyme Test. It was found that the more intelligible speaker had CVAR's 2-4 dB higher than the less intelligible speaker. Hecker (1974) attempted to increase speaker intelligibility by increasing the CVAR. This was accomplished by splicing out the consonant, increasing its amplitude, and re-splicing it into the word. After processing, intelligibility based on the Modified Rhyme Test showed an increase from 78% to 81% at 4dB of signal-to-noise ratio, a small but significant increase. A number of studies have also considered changes in vocal effort (Perkell and Klatt, 1986) and presence of the Lombard effect (Hanley and Harvey, 1965; Pearsons, et al., 1977; Junqua, 1993,96).

Below, we summarize the primary findings from analysis of mean and variance in word and phoneme class intensity for various speech styles under stress.

Summary of Key Results: Mean Intensity vs. Stress

Mean RMS intensity from highest to lowest versus speaking style for all speech classes (* = significant with respect to neutral)

Shift of Mean RMS Intensity

Word		Vowel		Consonant		Semivowel		Diphthong	
A*	+48%	A*	+32%	SO	+33%	A	+16%	L*	+46%
L*	+38%	L*	+25%	C7	+23%	SO	0%	A*	+45%
LM	+8%	C	+2%	C5	+14%	F	0%	LM	+8%
SL	+4%	LM	+1%	A	+12%	N		C	+3%
F	+2%	SL	+1%	SL	+3%	L	-6%	F	+3%
N		N		F	+2%	SL	-7%	N	
SO	-5%	F	-2%	LM	+1%	C5	-15%	SL	-1%
C	-8%	SO	-3%	N		C7	-17%	C5	-3%
C5	-8%	C7	-6%	C	-8%	LM	-17%	C7	-4%
C7*	-10%	C5	-8%	L	-17%	C	-18%	SO	-7%

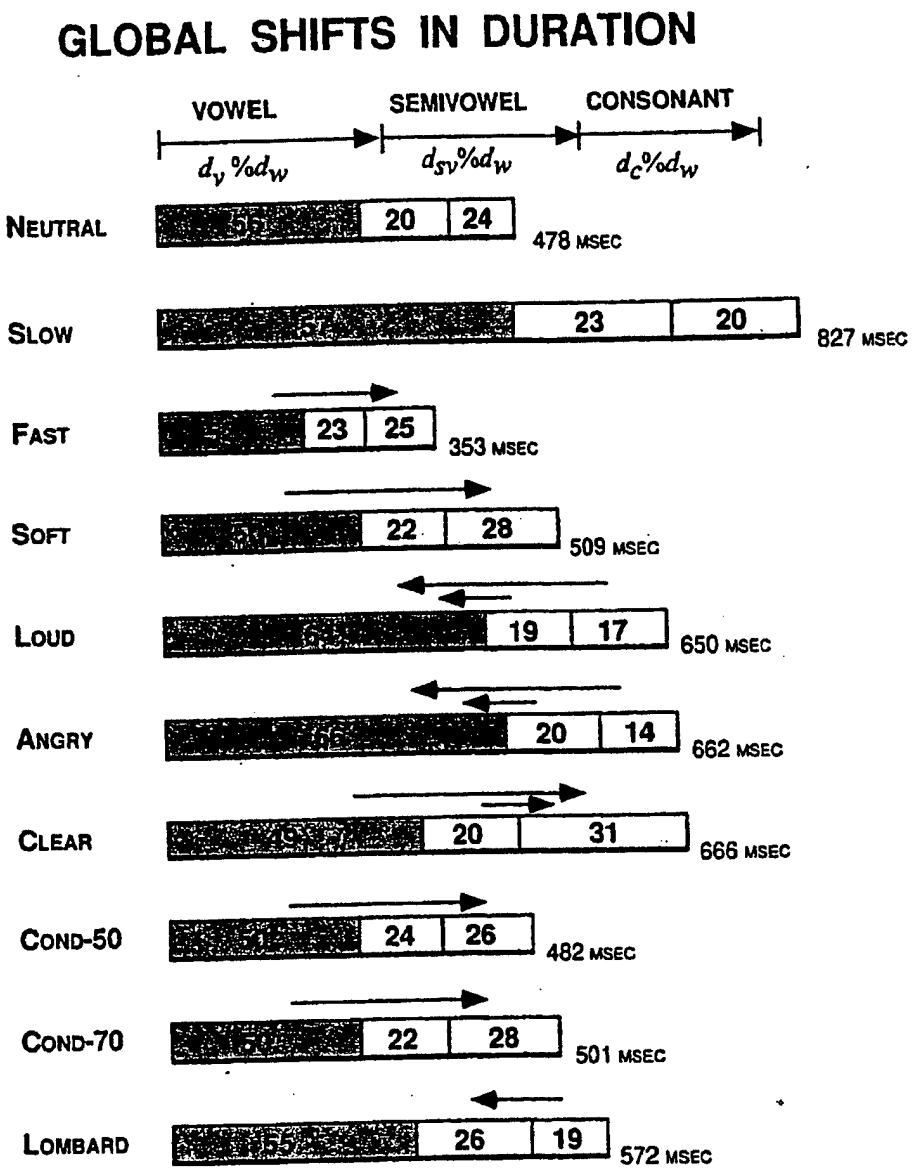


Figure 2: A pictorial representation of global duration shifts for speech under stress. The length of each bar graph is proportional to each style's average duration. Speech class percentages shown for each style are based on an ideal word containing one phoneme from each class. Arrows indicate significant shifts in duration based on phoneme ratios as a result of stress.

- Mean RMS word intensity values may be used as significant indicators for speech in angry, loud, and high workload task styles when compared to neutral conditions
- Loud and angry mean RMS word intensity are significantly different from all other styles considered
- Loud and angry mean RMS vowel and diphthong intensities are significantly different from all styles considered
- Mean RMS consonant and semivowel intensity are not significant stress indicators for any style considered
- Mean RMS intensity is not a significant indicator for moderate versus high task workload conditions

Summary of Key Results: Intensity Variance vs. Stress

- RMS intensity variance from highest to lowest versus stress style
  (* = significant with respect to neutral)

Shift in the Variance of RMS Intensity

Word		Vowel		Consonant		Semivowel		Diphthong	
A*	+312%	A*	+107%	SO*	+64%	A*	+346%	N	
L*	+154%	L	+25%	A	+54%	L	+80%	C7	-20%
LM	+37%	C	+17%	C7	+52%	C7	+46%	F	-25%
SO	+30%	C7	+8%	SL	+47%	C5	+27%	C	-39%
N		F	0%	L	+19%	LM	+19%	C5	-40%
F	-18%	N		C	+14%	F	+6%	SO	-43%
C5	-25%	C5	-1%	C5	+11%	N		A	-50%
SL	-31%	LM	-14%	LM	+10%	SL	-17%	L	-55%
C7	-33%	SL	-18%	N		SO	-20%	SL	-62%
C	-47%	SO	-32%	F	-8%	C	-41%	LM	-78%

- Variance of RMS word intensity may be used as a significant indicator for speech in angry and loud styles when compared to neutral
- Variance of loud and angry RMS word intensity is significantly different from most other styles considered
- Variance of angry RMS vowel and semivowel intensities were significantly different from most styles considered
- Variance of RMS consonant and diphthong intensity were not significant stress indicators for most styles
- Variance of RMS intensity (for word or phoneme class) was not a significant indicator for moderate versus high workload task conditions

Next, it may be beneficial to reflect on the intensity variation between individual phoneme classes. Consider the case when a talker is speaking under fast or Lombard effect in noisy environmental conditions. A talker could maintain overall word intensity, yet vary a particular phoneme class with respect to another. Hence, several average RMS ratio measures were

formed. Three ratios were considered: i) consonant versus vowel amplitude ratio (CVAR), ii) consonant versus semivowel amplitude ratio (CSVAR), and iii) vowel versus semivowel amplitude ratio (VSVAR). These ratios are used to determine in which directions speakers vary their intensity patterns under stress.

CVAR's for fast, slow, and Lombard effect conditions were relatively constant. Increased CVARs resulted for soft and both task conditions, which suggest talkers emphasize consonant amplitude with respect to vowel amplitude under these stress conditions. Decreases in CVAR's for loud, angry, and clear styles signify further importance in vowel rather than consonant amplitudes. CSVARs also demonstrate a talker's emphasis of consonant versus semivowel amplitudes for soft, Lombard, and both task conditions. Decreased CSVAR was noted for only loud and angry styles. Finally, VSVAR generally result in further vowel emphasis. Only the soft speaking style results in decreased VSVAR, with loud having the highest. In order to get an overall perspective of changes in intensity across the various styles, a pictorial representation is shown in Fig.3. The clear bar graphs are proportional to RMS word intensity for each style. Shaded regions within each bar graph indicate average RMS intensity values for vowel, semivowel, and consonant phoneme classes[6]. Triangles below each bar graph indicate statistically significant shifts in phoneme class intensity. A single arrow indicates a strong shift in average RMS intensity from one class to the other. A double arrow indicates an extreme shift in RMS intensity (from weaker to stronger). Phoneme class shifts indicate opposite movement for soft and loud speech classes. For loud speech, vowel amplitudes are strongly emphasized, while in soft speech consonant amplitudes are emphasized. Fast speech had little or no movement between phoneme classes.

[6] Phone class heights have all been scaled by a factor of 2/3 so word RMS intensities are visible in the presentation.

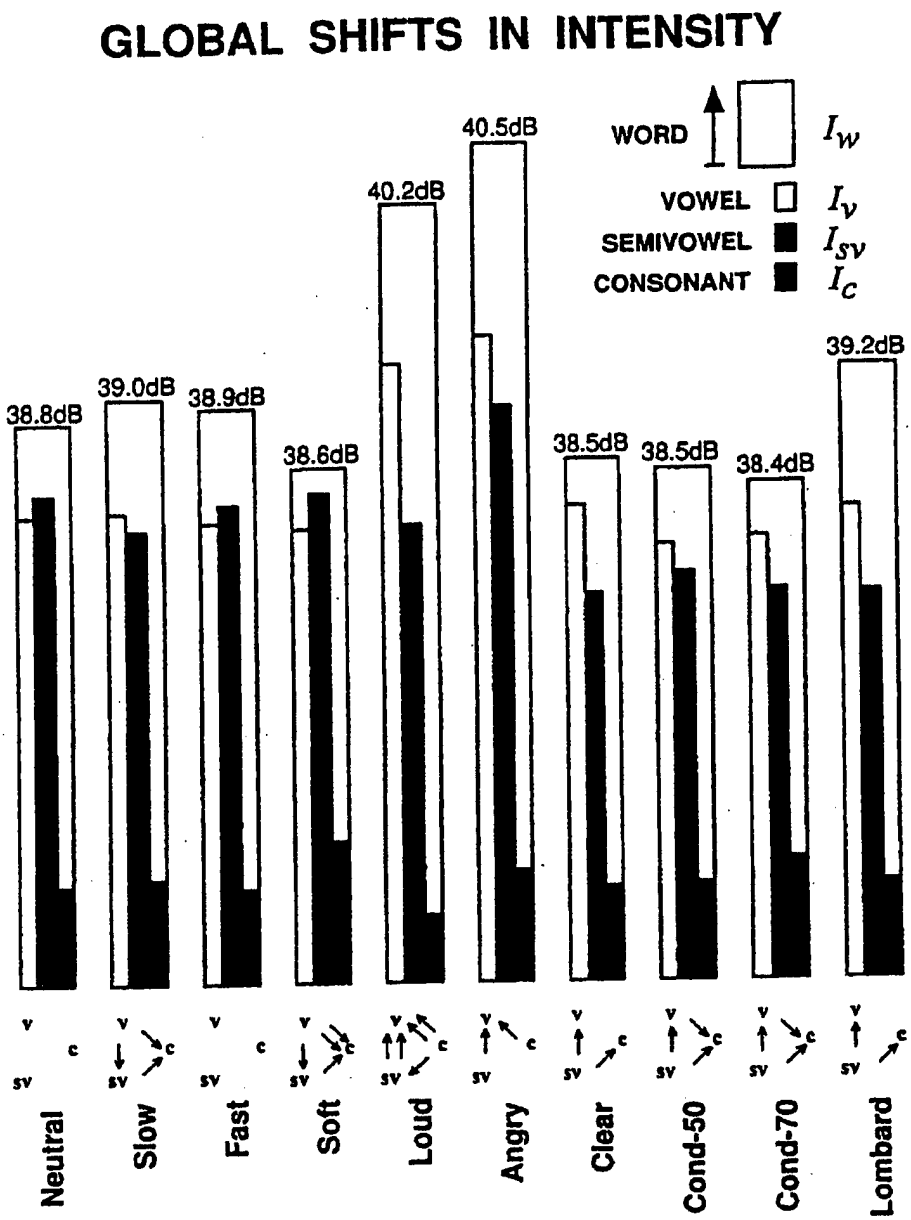


Figure 3: A pictorial representation of global intensity and shifts in intensity for speech under stress. The height of each bar graph is proportional to each style's average RMS word intensity. Speech class RMS values are also shown. Arrows indicate significant shifts in intensity as a result of stress.

3.5 Glottal Source Spectral Analysis

In this section, we focus on glottal source effects. There are a variety of characteristics relating to speech excitation which are adjusted to convey the stress or emotional state of a speaker. We have seen that the fundamental frequency of vocal fold movement is a statistically reliable indicator of many stressful speaking conditions. In addition to rate, aspects such as duration of each laryngeal pulse (both open and closed glottal periods), the instant of glottal closure, or the shape of each pulse play important roles in a talker's ability to vary source characteristics.

An analysis of glottal source spectral characteristics was performed for SUSAS speech data. Utterances rich in vowel content but lacking adjacent nasal portions were chosen. An algorithm was developed for analysis of the distribution of frame energy for each stress condition. The division between voiced (high frame energy) and unvoiced (low frame energy) speech is quite apparent in all cases. These frame energy distributions can also serve as possible stress indicators. For example, high energy frame concentration increased for angry, loud, and Lombard conditions. However, low energy frame concentration increased for clear, slow and soft conditions. A shift was also observed for frames with moderate energy (40 to 60 dB) toward primarily higher regions. This was observed for loud and angry styles, thus indicating that under these conditions, the time duration spent during transitional periods between voiced and unvoiced portions is reduced. A voiced energy cutoff was selected from corresponding frame energy distributions close to the upper peak in the frame distributions (i.e., normally between 65 to 70 dB). Frames above the threshold are extracted and a gain normalized periodogram spectral estimate found for each frame. Periodograms from all selected frames are averaged to remove the effects of the varying vocal-tract response. This leaves an estimate of the glottal source spectrum. Each selected frame's energy is also averaged to obtain a final gain factor for the glottal spectrum. This was performed for each of the ten stress conditions.

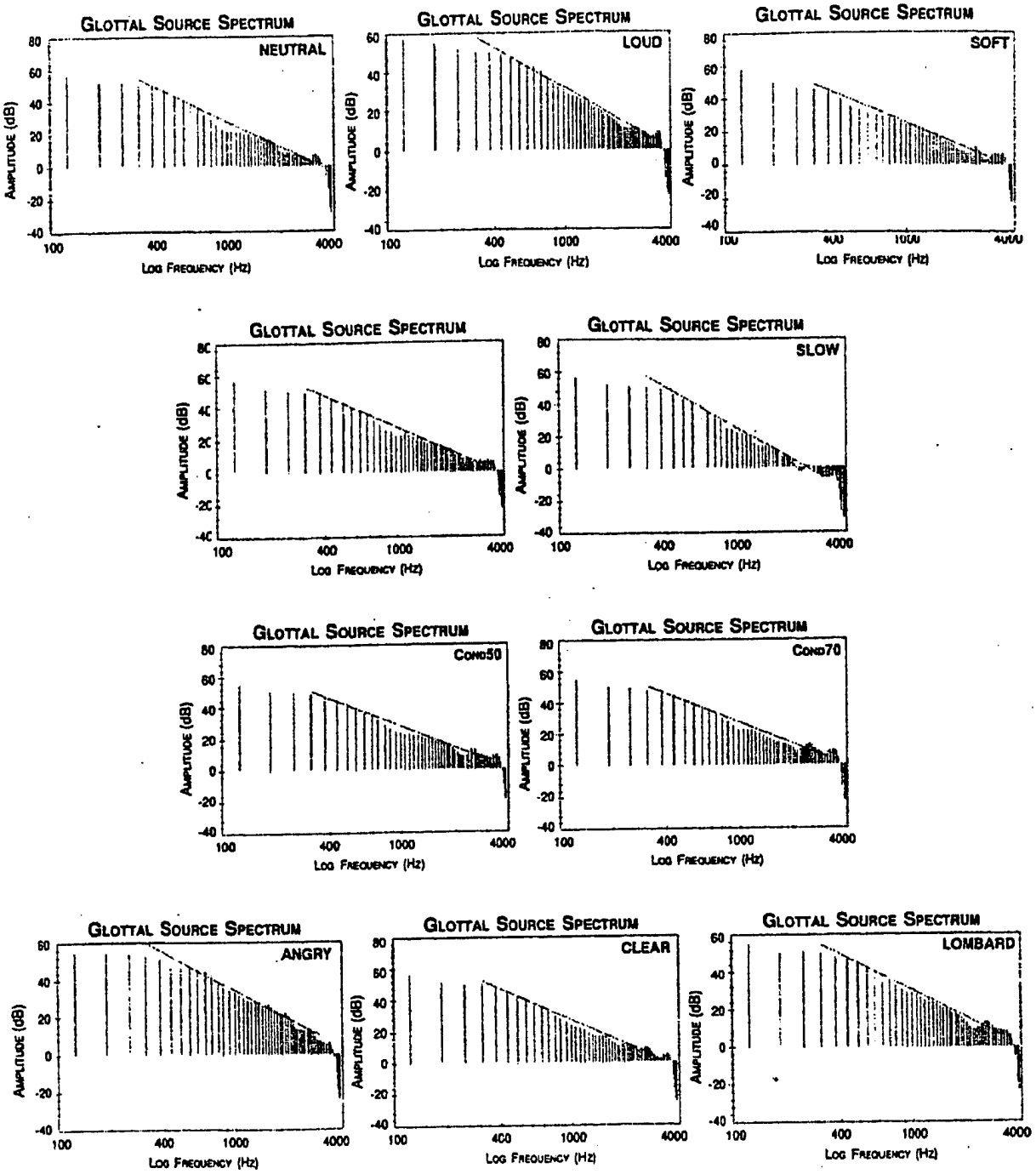


Figure 4: Glottal source spectral estimates based on non-nasalized utterances for neutral, loud, soft, fast, slow, moderate (Cond50\%) and high (Cond70\%) computer workload conditions, angry, clear, and Lombard effect stress talking styles. (Note: A +6dB/octave spectral roll-off should be added to remove the effects of the lip radiation component.)

3.6 Analysis of Vocal Tract Articulatory Characteristics

It is reasonable to hypothesize that stress factors will also affect the position and rate of change of the articulators, which shape the vocal-tract. It has been suggested that these changes may represent a major contributor to the reduced performance of present-day recognition algorithms in stressful environments (Hansen, 1988, 1996). Therefore, we consider an initial analysis of vocal tract structure based on sample articulatory profiles. Previous articulatory studies have considered methods to estimate vocal tract configuration based on the acoustic signal (Kobayashi, Yagyu, Shirai, 1991; Wakita, 1973). The analysis here is based upon a linear acoustic tube model with speech sampled at 8 kHz. In order visualize the effects of stress on physical vocal tract shape, the movements throughout the vocal tract can be displayed by superimposing a time sequence of estimated vocal tract shapes for a chosen phoneme. The vocal tract shape analysis algorithm assumes a known normalized area function and acoustic tube length. The algorithm begins by computing the sagittal distance function by assuming a cylindrical vocal tract. Next, a set of rigid points from the glottis to the upper teeth (and rigid upper lip) models the hard palate. With the hard palate model in place, the soft palate and pharynx are approximated by forming a dependence upon the sagittal distance function. Finally, the lower lips are modeled using one of four rigid models dependent upon the acoustic tube length.

An analysis of articulatory changes in vocal tract shape under neutral and various stress conditions was performed for extracted phoneme sequences from SUSAS. Fig.5 illustrates a set of vocal tract shapes which are superimposed for each frame in the analysis window (the number of extracted frames are summarized for each stress condition). For *Neutral*, there is some movement of the articulators in the pharynx and oral cavities (as there should be for the production of the /r-iy/ phone sequence in "freeze"). There is also limited movement for the *Soft* speaking condition. However, for *Angry*, *Loud*, and *Lombard* conditions, there is significant perturbation in the blade and dorsum of the tongue and the lips. This extreme vocal tract variation is also present in the same phone sequence from the *Actual* stress domain (speech from roller-coaster rides). This suggests that when a speaker is under stress, typical vocal tract movement is effected, suggesting a quantifiable perturbation in articulator position. This

characteristic was used as one of several features for a study on stressed speech classification by Womack and Hansen (1996).

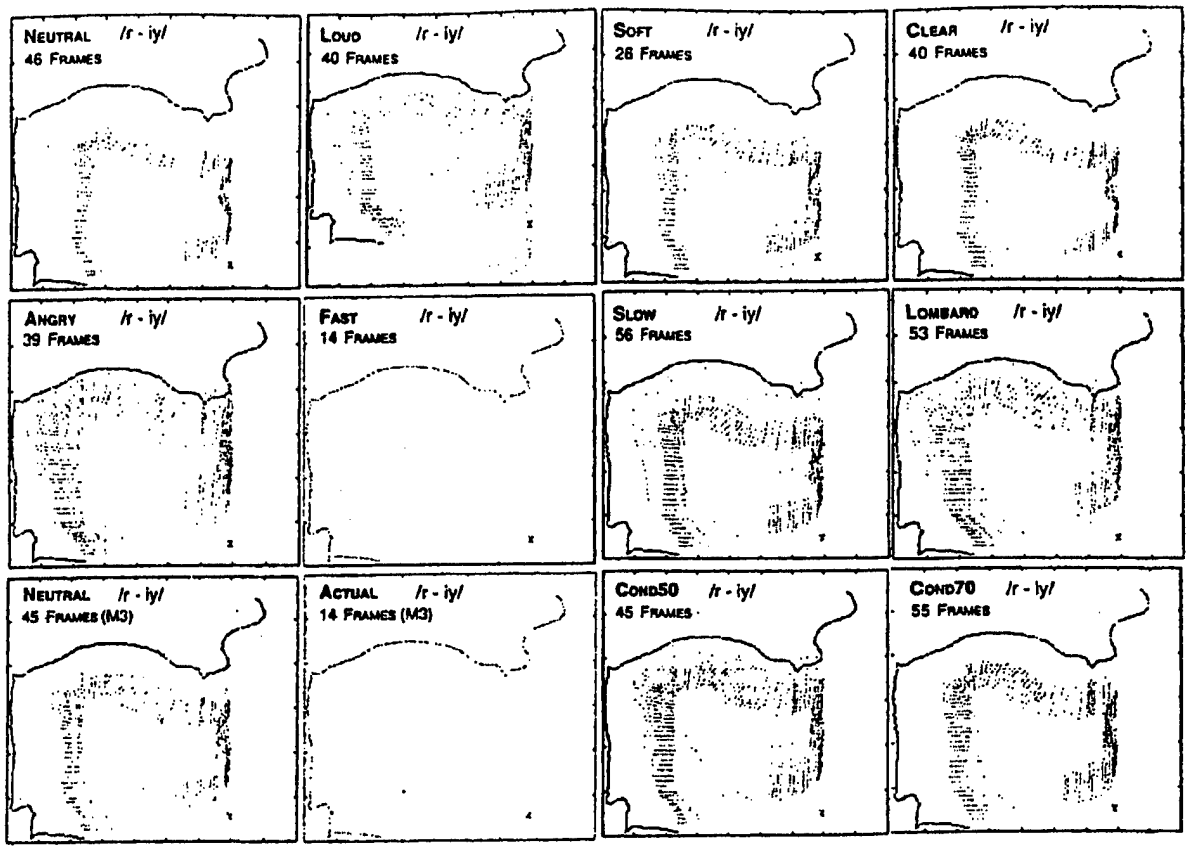


Figure 5: Sample vocal tract articulatory profiles for the phoneme sequence /r-iy/ from the word *freeze* across SUSAS speech under stress conditions. Each speech frame analysis width was 24ms, with a skip rate of 8ms. The number of frames indicate the /r-iy/ duration over which the profiles are plotted (e.g., 39 frames x/ 8ms/frame = 312ms).

3.7 Vocal Tract Spectrographic Analysis

Our initial analysis of vocal tract spectral structure focuses on sample spectrographic analysis of speech under stress. Several hundred spectrograms were informally compared across stress conditions in SUSAS. Fig.6 illustrates example responses for the word *help* spoken under stress conditions. Final stop releases were in general not present in high stress styles such as *Angry*,

Loud, and most *Actual* stress examples. Stop release time was normally longer for *Clear* and at times *Lombard effect* conditions. For *Angry, Loud*, and *Lombard* stress conditions, the high frequency energy generally increases with more irregular formant structure. Formants are also higher in amplitude and more clearly defined. This was partially confirmed in the previous analysis on the glottal source spectrum.

The spectral characteristics illustrated in spectrographic analysis suggest that the presence of stress based information can be obtained from a statistical analysis of formant location and bandwidth. A more complete discussion of the statistical analysis of individual phonemes across formant location and bandwidth can be found in Hansen, 1998b.

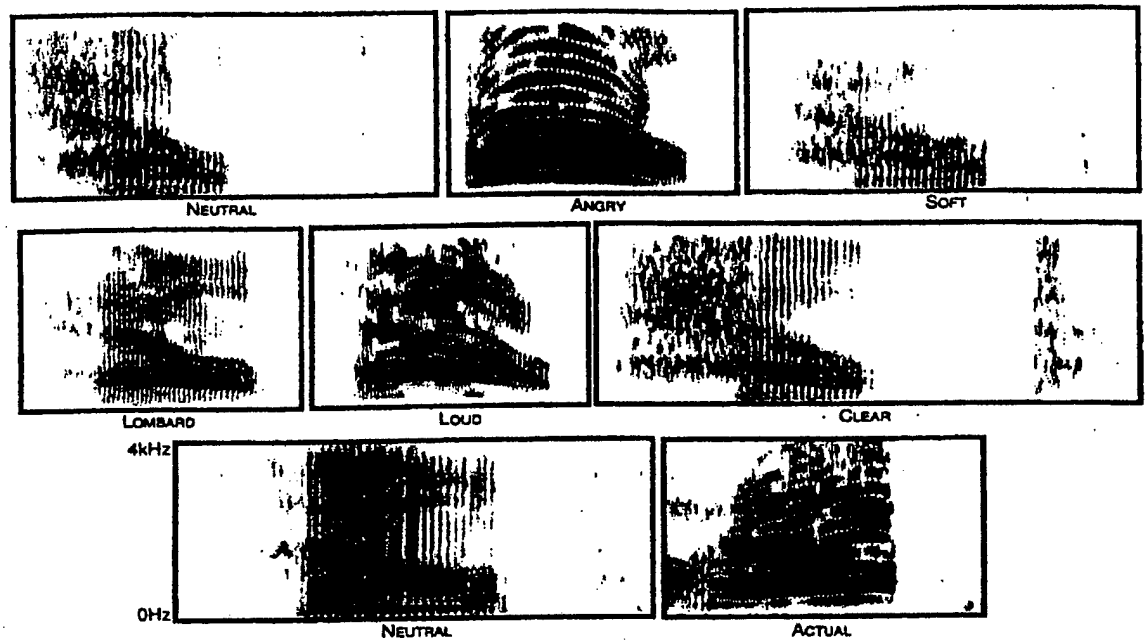


Figure 6: Spectral Responses under Stress Caption Sample vocal tract spectral responses from *help* utterances in the SUSAS speech under stress database.

4 Stress Classification

The field of computer based voice stress detection is an emerging area. There has been some activity in commercial voice stress analysis in applications for forensic science (Cestaro, 1995). These methods are typically based on some aspects of pitch perturbation or micro-tremors (Lippoid, 1971). However, these commercial systems are not universally accepted by speech scientists because the 'stress' in these cases is normally associated with deception.

The focus here is exclusively on stress based speech production variations resulting from task workload, Lombard effect, or emotional/psychological changes. Recently, a number of studies have been reported which focus directly on the formulation of computer based stress detection.

Several methods have been proposed based on neural network classifiers. For example, Hansen and Womack (1996) considered a neural network based stress classifier using five different cepstral feature sets. Features that were found to be the most useful were the auto-correlation Mel AC_i and cross-correlation Mel $XC_{i,j}$ (XC-Mel) cepstral parameters. Further classification studies have expanded on these neural network approaches using target driven features (Womack and Hansen, 1996). In that method, a wide selection of features were automatically extracted including articulatory measures, pitch, phone duration, and a variety of spectral based information. Next, the most effective feature subset for each targeted stress condition was determined during training, and only those targeted features used during classification. This allows the classifier to use the most discriminating features for classification of each stress style.

Other methods have also been proposed based on the nonlinear Teager Energy Operator (TEO) (Cairns and Hansen, 1994a, 1994b), where the shape of a duration normalized energy profile was used in a hidden Markov model based stress classifier. That study, clearly demonstrated the potential a TEO-based feature could have in improving stress classification performance. Motivated by this work, several recent investigations have considered more extensive feature processing methods based on TEO principles (Zhou, Hansen, Kaiser 1998a,b).

In the following sections, we first consider stress classification experiments which use linear based speech features and optimum Bayesian detection theory. These experiments were

conducted on features such as duration, intensity, pitch, glottal source, and vocal tract spectrum. In a recent study, several TEO-based nonlinear features were found to be very effective for stress classification (Zhou, Hansen, Kaiser 1998). Therefore, in Section 4.3 one nonlinear based feature is described, with results presented for both classification and assessment of speaker stress.

4.1 Bayesian Stress Classification with Linear Speech Features

Having established relationships between speech production under stress and speech feature variation (Hansen 1998a,b), we now turn to the related problem of classification of speech under stress. Our task here is to formulate an algorithm for detection of speech spoken under one particular stress style versus neutral speech. It has been shown that there are observable differences in duration, intensity, pitch, glottal source information, and formant locations between neutral and stressed speech. Therefore, it is worthwhile to evaluate their performance for stress classification, or stress detection. Here the two terms, classification and detection, can be used interchangeably since only pairwise classification is considered. Two processing stages are required for stress detection. In the first stage, acoustical features are extracted from an input speech waveform. The second stage is focused on detection of stressed speech from neutral using one or more available methods. A variety of methods exist for stress detection which include, but not limit to, detection-theory based methods, methods based on distance measures, neural network classifiers, and statistical modeling based techniques. In this section, we employ two methods, one using a Bayesian hypothesis-testing framework, and the other using a distance measure to detect stressed versus neutral speech.

4.1.1 Description of Features

For the five linear features used for stress classification, only vowel sections were extracted from the simulated domain of the SUSAS database for evaluation. The sample length of each vowel in msec is used as the duration feature. The intensity feature is defined as,

$$Intens = \sqrt{\frac{1}{K}\sum_{i=1}^{k} s^2(i)} \tag{9}$$

where $s(i)\,(i=1,\cdots,K)$ represents the K individual samples in the vowel. Pitch, glottal source information, and formant locations are extracted on a frame basis with frame length being 32 msec and an overlap length between adjacent frames of 16 msec. The modified simple inverse filter tracking (MSIFT) algorithm (Arslan, 1996) is employed to extract pitch frequencies from vowel speech waveforms. Spectral slope was used as the glottal source feature. It is difficult to obtain the glottal spectral slope from the raw vowel speech waveform due to the coupling effect between the sub-glottal structure and forward portion of the vocal tract. To avoid this effect, only data obtained during closed vocal fold periods was used, which unfortunately limits the available data. Also, it is difficult to accurately locate the boundaries between vocal fold closing and opening periods. As an approximation, a frame based log average amplitude FFT was computed versus log frequency for each vowel section and used to determine boundaries.

The fourth feature is the slope of the glottal source spectrum. A straight line is used to approximate excitation spectral envelope, and the line's slope is considered as the glottal spectral slope. Finally, for the last set of features, the first two formant locations are used, since these were shown shown to change measurably between neutral and stressed speech (Hansen 1998b). Here, the ESPS/xwaves function "formant" was employed to extract formant locations for all vowels in the SUSAS database (Entropic, 1993).

4.1.2 Detection-theory Bayesian Hypothesis Testing

A flexible framework for stress detection can be easily established using detection theory. For such a scheme, there are two hypotheses termed $H0$ and $H1$. Under $H0$, the speech is neutral; while under $H1$, the speech is stressed. Given an input speech feature vector, $x, (x=\chi_1,\cdots,\chi_M;$ M is the vector length) the following two conditional probability densities (PDF) are estimated, $p(x|H0)$ and $p(x|H1)$. With these PDFs, the likelihood ratio, λ, is then defined as,

$$\lambda = \frac{p(x|H1)}{p(x|H0)}. \tag{10}$$

The decision of whether the input speech is neutral or stressed is made by comparing the likelihood or log likelihood ratio with a pre-defined threshold, β. If it is bigger than β, the input

speech is labeled as stressed; otherwise it is classified as neutral. The value of β depends on what criterion is used for detection. In a stress classification system, a criterion should be selected so that the two important probabilities, the false acceptance rate (FAR) and the false rejection rate (FRR), should be as low as possible. Obviously, it is not possible to minimize both FAR and FRR, and hence, a compromise must be made between FA and FR. For some systems, the requirement for one probability is more important than the other. For a stress classification system, however, we are only interested in the overall accuracy and have no preference for either FAR or FRR. Therefore, the value of β corresponding to equal error (FAR=FRR) rate (EER) is selected. In the experiments performed here, the values of FAR and FRR were calculated as the ratio of the number of falsely accepted vowels to the total number of vowels, and the ratio of the number falsely rejected vowels to the total number of vowels, respectively. By changing the threshold value, the value of β corresponding to EER can be found.

In order to form the likelihood ratio in Eq.10, we must first estimate the PDFs $(p(x|H0)\text{and}\, p(x|H1))$ of both the neutral and stressed speech features. If we assume that all components $(\chi_1, \chi_2, \cdots, \chi_M)$ of the feature vector x are independent and identically distributed Gaussian random variables with mean, μ_n, and variance, σ_n^2, under neutral conditions, but with a different mean, μ_s and variance σ_s^2 under stressed conditions, then the individual feature component PDFs conditioned on neutral $(H0)$ or stressed $(H1)$ speech is as follows,

$$f(\chi_i|H0) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(\chi_i - \mu_n)^2}{2\sigma_n^2}\right), \tag{11}$$

$$f(\chi_i|H1) = \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left(-\frac{(\chi_i - \mu_n)^2}{2\sigma_s^2}\right). \tag{12}$$

With these PDFs and assuming statistical independence, the overall conditional probabilities $p(x|H0)\text{and}\, p(x|H1)$ can be computed as,

$$p(x|H0) = \left(2\pi\sigma_n^2\right)^{-\frac{M}{2}} \exp\left(-\frac{1}{2\sigma_n^2}\sum_{i-1}^{M}(\chi_i - \mu_n)^2\right), \tag{13}$$

$$p(x|H1) = \left(2\pi\sigma_s^2\right)^{-\frac{M}{2}} \exp\left(-\frac{1}{2\sigma_s^2}\sum_{i=1}^{M}(\chi_i - \mu_s)^2\right). \tag{14}$$

Substituting Eq.13 and 14 into Eq.10, the likelihood ratio can be computed as,

$$\lambda = \frac{p(x|H1)}{p(x|H0)} \tag{15}$$

$$= \left(\frac{\sigma_n}{\sigma_s}\right)^M \exp\left(\frac{1}{2\sigma_n^2}\sum_{i=1}^{M}(\chi_i - \mu_n)^2 - \frac{1}{2\sigma_s^2}\sum_{i=1}^{M}(\chi_i - \mu_s)^2\right)$$

Taking the logorithm of each side, the log likelihood ratio is obtained as follows,

$$\ln\lambda = M\ln\left(\frac{\sigma_n}{\sigma_s}\right) + \frac{1}{2\sigma_n^2}\sum_{i=1}^{M}(\chi_i - \mu_n)^2 - \frac{1}{2\sigma_s^2}\sum_{i=1}^{M}(\chi_i - \mu_s)^2, \tag{16}$$

$$= M\ln\left(\frac{\sigma_n}{\sigma_s}\right) + \frac{1}{2\sigma_n^2}\sum_{i=1}^{M}(\chi_i - \hat{\mu} + \hat{\mu} - \mu_n)^2 - \frac{1}{2\sigma_s^2}\sum_{i=1}^{M}(\chi_i - \hat{\mu} + \hat{\mu} - \mu_s)^2,$$

$$= M\ln\left(\frac{\sigma_n}{\sigma_s}\right) + \frac{1}{2\sigma_n^2}\sum_{i=1}^{M}(\chi_i - \hat{\mu})^2 + \frac{M}{2\sigma_n^2}(\hat{\mu} - \mu_n)^2 - \frac{1}{2\sigma_s^2}\sum_{i=1}^{M}(\chi_i - \hat{\mu})^2 - \frac{M}{2\sigma_s^2}(\hat{\mu} - \mu_s)^2,$$

$$= M\ln\left(\frac{\sigma_n}{\sigma_s}\right) + \frac{M}{2\sigma_n^2}\left(\hat{\sigma}^2 + (\hat{\mu} - \mu_n)^2\right) - \frac{M}{2\sigma_s^2}\left(\hat{\sigma}^2 + (\hat{\mu} - \mu_s)^2\right),$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ and are the estimated mean and variance of the input sample feature vector, x, which are defined as,

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M}\chi_i, \tag{17}$$

$$\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(\chi_i - \hat{\mu})^2. \tag{18}$$

Similarly, if we assume that all feature vector components $(\chi_1, \chi_2, \cdots, \chi_M)$ are independent and identically distributed from a Gamma distribution, $\Gamma(a, \beta)$ with $a = a_n$ and $\beta = \beta_n$ under neutral conditions, but with $a = a_s$ and $\beta = \beta_s$ under stressed conditions, the PDFs are formed as,

$$f(\chi_i|H0) = \begin{cases} \frac{\beta_n^{-a_n}}{\Gamma(a_n)} \chi_i^{(a_n-1)} e^{-\chi_i/\beta_n}, & \chi_i > 0 \\ 0, & \chi_i \le 0 \end{cases} \tag{19}$$

$$f(\chi_i|H1) = \begin{cases} \frac{\beta_s^{-a_s}}{\Gamma(a_s)} \chi_i^{(a_s-1)} e^{-\chi_i/\beta_s}, & \chi_i > 0 \\ 0, & \chi_i \le 0 \end{cases} \tag{20}$$

The conditional probabilities are then obtained as,

$$p(x|H0) = \begin{cases} \left(\frac{\beta_n^{-\alpha_n}}{\Gamma(\alpha_n)}\right)^M \left(\prod_{i=1}^{M} \chi_i\right)^{(\alpha_n-1)} \exp\left(-\frac{1}{\beta_n}\sum_{i=1}^{M}\chi_i\right) & \chi_i > 0, i=1,2\cdots,M \\ 0, & \text{, otherwise} \end{cases} \tag{21}$$

$$p(x|H1) = \begin{cases} \left(\frac{\beta_S^{-\alpha_s}}{\Gamma(\alpha_s)}\right)^M \left(\prod_{i=1}^{M} \chi_i\right)^{(\alpha_s-1)} \exp\left(-\frac{1}{\beta_s}\sum_{i=1}^{M}\chi_i\right) & \chi_i > 0, i=1,2\cdots,M \\ 0, & \text{, otherwise} \end{cases} \tag{22}$$

Substituting Eq.21 and 22 into Eq. 10 we obtain the likelihood ratio, λ, and the log likelihood ratio, $\ln\lambda$, for the case where sample features are Gamma distributed $(\chi_i > 0, i = 1,2\cdots,M)$, as follows,

$$\begin{aligned} \lambda &= \frac{p(x|H1)}{p(x|H0)} \\ &= \left(\frac{\beta_s^{-\alpha_s}\Gamma(\alpha_n)}{\beta_n^{-\alpha_n}\Gamma(\alpha_s)}\right)^M \left(\prod_{i=1}^{M}\chi_i\right)^{(\alpha_s-\alpha_n)} \exp\left(M\hat{\mu}\left(\frac{1}{\beta_n}-\frac{1}{\beta_s}\right)\right), \end{aligned} \tag{23}$$

$$\ln\lambda = \left(\frac{\beta_s^{-\alpha_s}\Gamma(\alpha_n)}{\beta_n^{-\alpha_n}\Gamma(\alpha_s)}\right) + (\alpha_s - \alpha_n)M\hat{\mu}_{1n} + M\hat{\mu}\left(\frac{1}{\beta_n}-\frac{1}{\beta_s}\right), \tag{24}$$

where $\hat{\mu}$ is the estimated mean of the input sample vector, **x**, as defined by Eq. 17, and $\hat{\mu}_{\ln}$ is defined as,

$$\hat{\mu}_{\ln} = \frac{1}{M}\sum_{i=1}^{M} \ln \chi_i \tag{25}$$

The decision of whether the input speech is neutral or stressed is made by comparing the likelihood (Eq. 15 for Gaussian distributed features or 23 for Gamma distributed features) or log likelihood ratio (Eq. 16 or 24) with a pre-defined threshold, β. If it is bigger than β, the input speech is labeled as stressed; otherwise it is classified as neutral. The value of β depends on what criterion is used for detection. In a stress classification system, a criterion should be selected so that the two important probabilities, the false acceptance rate (FAR) and the false rejection rate (FRR), should be as low as possible. Obviously, it is not possible to minimize both FAR and FRR, and hence, a compromise must be made between FA and FR. For some systems, the requirement for one probability is more important than the other. For a stress classification system, however, we are only interested in the overall accuracy and have no preference for either FAR or FRR. Therefore, the value of β corresponding to equal error (FAR=FRR) rate (EER) is selected. In the experiments performed here, the values of FAR and FRR were calculated as the ratio of the number of falsely accepted vowels to the total number of vowels, and the ratio of the number falsely rejected vowels to the total number of vowels, respectively. By changing the threshold value, the value of β corresponding to EER can be found.

4.1.3 Distance Measure Testing

It is also possible to detect stressed speech from neutral using a distance measure with prior trained feature distributions. Given an input speech feature vector, ($\mathbf{x}=\chi_1,\chi_2,\cdots,\chi_M$); M is the vector length, two values, the distance between **x** and the neutral speech feature distribution, d_n, and the distance between **x** and the stressed speech feature distribution, d_s, are computed as,

$$d_n = \frac{|\hat{\mu} - \mu_n|}{\hat{\sigma}\sigma_n}, \tag{26}$$

$$d_s = \frac{|\hat{\mu} - \mu_s|}{\hat{\sigma}\sigma_s}, \tag{27}$$

where $\mu_n, \sigma_n, \mu_s, \sigma_s$ are means and standard deviations for the neutral and stressed speech features, which are obtained from training data; $\hat{\mu}$ and $\hat{\sigma}$ are the sampled estimated mean and standard deviation of the components of the input vector, **x**, as defined in Eq.17 and 18 respectively. This distance measure reflects how close the input test speech feature vector is to the feature distributions of neutral and stressed speech data. If d_n is smaller than d_s, the input vector **x** is labeled as neutral, otherwise, it is assigned as stressed. The distance scores can also be used to quantify the degree of stress content in the test data.

4.2 Linear Feature Based Evaluations

A 33 word vocabulary[7] under neutral, angry, loud, and Lombard effect speaking styles from the simulated domain of the SUSAS database was employed for evaluations. For each test token, all samples corresponding to vowels were extracted. Other voiced data such as diphthongs, liquids, glides, and nasals were not considered due to changing spectral structure from articulatory movement. It is believed that the muscle control needed for such articulatory movement would also be effected under stress. Vowels were selected to investigate vocal fold changes under stress and static vocal tract adjustments due to stress. From all identified vowels, duration, intensity, pitch, glottal spectral slope, and formant locations were extracted. For each feature, all extracted data was used to estimate the density function of the feature distribution, and then obtain the ROC (receive operation characteristic) curve for the Bayesian hypothesis-testing method. In order to achieve open-set performance in the test phase, the entire vowel data set was first divided for each feature into 10 equal-size groups. For each set of the 10 groups, one group is set aside and the remaining data (9 groups) used to obtain the EER threshold for Bayesian

[7] For these evaluations, the two words "destination" and "histogram" were set aside because of the increased impact of lexical stress on polysyllable words. The remaining 33 word vocabulary consisting of 26 monosyllable words and 7 two-syllable words.

hypothesis-testing method, and the mean and variance for the distance measure approach. The final error rate is obtained by accumulating all error rates from 10 open-set tests. The next five subsections consider stress detection performance using the optimum Bayesian detection scheme for a feature in each speech production domain (duration, intensity, pitch, glottal source, vocal tract spectrum).

4.2.1 Duration

For the Bayesian hypothesis-testing method, PDFs of vowel duration were first estimated to form the likelihood ratio. Using phone segment label information, a vowel duration histogram was obtained, following by fitting a Gamma distribution to the data histogram (examples for the loud speaking style are shown in Fig. 7a. Based on this Gamma pdf, the ROC for open-set test performance was obtained for the Bayesian hypothesis-testing method. To find average test results, the data was divided for each feature into 10 equal size sets. For each of the 10 sets, we test with one set and train with the other 9 to calculate the average EER threshold for the Bayesian hypothesis testing approach, and the mean and variance of the feature distribution for the distance measure approach. Fig.8a shows the ROC of detecting speech under "loud" speaking style from neutral speech using duration. Table 3 lists the open-set test results by using the Bayesian hypothesis-testing method as well as using the distance measure approach. Several testing feature vector lengths (1, 5, 10) were used to obtain ROC curves and error rates. From the results in ROC and table, increasing the input vector length does not significantly improve the detection accuracy (especially for detection of Lombard effect versus neutral speech using the Bayesian hypothesis-testing method). Also, Table 3 shows that the distance measure approach produces slightly better performance for Lombard effect and loud speech, but slightly lower results for angry speech when compared with the Bayesian hypothesis-testing method. In general, given the error rate levels for the three stress classes tested, vowel duration is not a strong feature for stress detection.

Table 3: Error Rate (percentage) of open-set Stress Detection Test using Duration as the feature.

Detection Method	Vector Length	**Error Rate:** Stress Style of Test Speech **DURATION**					
		Neutral	Angry	Neutral	Loud	Neutral	Lombard
Optimal Detection	1	45.13	45.38	38.21	38.72	40.77	40.26
	5	36.36	38.96	33.77	35.06	40.26	40.26
	10	41.03	35.90	38.46	35.90	38.46	46.15
Distance Measure	5	48.05	49.35	29.87	36.36	32.47	42.86
	10	43.59	53.85	28.21	41.03	30.77	46.15

4.2.2 Intensity

For each vowel, we use Eq. 9 to calculate its root mean square (RMS) intensity. From a plot of the histogram, it was determined that the Gaussian PDF would fit the intensity distribution well. From Eq. 9, it is clear that the intensity feature can never be negative while a Gaussian PDF ranges from $-\infty$ to ∞. To solve this conflict, a conditional PDF, $f(\chi|X\geq 0)$, is used to fit the intensity distribution. $f(\chi|X\geq 0)$ is obtained as follows,

$$f(\chi|X\geq 0)=\frac{1}{p_0\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\chi-\mu)^2}{2\sigma^2}\right), \tag{28}$$

$$p_0=1-\int_{-\infty}^{0}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\chi-\mu)^2}{2\sigma^2}\right), \tag{29}$$

where μ and σ^2 are the mean and variance. Fig. 7b shows how a conditional Gaussian PDF fits the intensity distribution for all vowels spoken under the loud stress condition.

Based on the conditional Gaussian PDF, the Bayesian hypothesis-testing method was used to classify stressed speech from neutral. ROC curves for each stress condition were obtained (sample ROC is shown in Fig. 8. In a similar manner to that used for duration, the open-set test results for the Bayesian hypothesis-testing method and distance measure approach were obtained and summarized in Table 4.

Table 4: Error Rate (percentage) of Open-set Stress Detection Test Using Intensity as the Feature

Detection Method	Vector Length	**Error Rate**: Stress Style of Test Speech **INTENSITY**					
		Neutral	Angry	Neutral	Loud	Neutral	Lombard
Optimal Detection	1	40.26	37.44	34.87	32.82	40.77	39.49
	5	24.68	22.08	27.27	22.08	38.96	35.06
	10	23.08	17.95	28.21	17.95	35.90	35.90
Distance Measure	5	41.56	27.27	35.06	22.08	40.26	35.06
	10	30.77	33.33	25.64	23.08	41.03	33.33

The ROC curves in Fig. 8 and open-set test results in Table 4 show that increasing input vector length does improve performance, especially for detecting angry and loud speech for the Bayesian hypothesis-testing method. As for the distance measure approach, increasing input vector length does not always improve performance. The open-set test results also show that both methods perform better for detection angry and loud speech than for detecting Lombard effect speech.

4.2.3 Pitch

Frame-based pitch measurements were extracted for the input neutral and stressed data, and the resulting histogram showed that a conditional Gaussian PDF was suitable for model distribution of this feature. Fig. 7c, Fig. 8c, and Table 5 show the pitch distribution, ROC curves for the Bayesian hypothesis-testing method, and open-set test results for both methods, respectively. Note that all zero pitch values are removed from ROC plots and open-set tests.

Table 5: Error Rate (percentage) of Open-set Stress Detection Test Using Pitch as the Feature

Detection Method	Vector Length	**Error Rate**: Stress Style of Test Speech **PITCH**					
		Neutral	Angry	Neutral	Loud	Neutral	Lombard
Optimal Detection	1	18.95	18.57	11.94	11.63	24.08	24.18
	5	15.17	14.31	10.34	10.00	21.90	22.07
	10	12.76	11.72	7.24	8.28	20.69	19.31
Distance Measure	5	15.34	15.00	12.41	7.07	23.10	19.48
	10	14.48	12.76	12.07	4.83	21.38	33.33

Compared to duration and intensity, pitch resulted in much better performance for stress detection. In a similar manner to intensity, pitch performs better for detection of angry and loud speech than for Lombard effect speech when using the Bayesian hypothesis-testing method. For detection of loud versus neutral speech, the Bayesian hypothesis-testing method achieves very

high accuracy. The distance measure method produced a similar level of performance with pitch as the feature.

4.2.4 Glottal Source Spectrum

For estimation of the glottal source spectral slope, only those vowels which were longer than 5 frames (i.e., 96 msec) are used (in order to get reliable slope estimates). Since glottal spectral slopes for vowel sections are almost all negative, the resulting feature histogram shows an envelope that is close to a Gamma distribution. In order to fit Gamma distribution to the feature histogram (shown in Eq. 19 and Eq. 20), only the absolute value of each spectral slope was considered (sample Gamma distribution for loud speech is shown in Fig. 7D. The ROC curves for the Bayesian hypothesis-testing method are shown in Fig. 8, and open-set test results for both Bayesian hypothesis-testing method and distance measure approach are summarized in Table 6.

Table 6: Error Rate (percentage) of Open-set Stress Detection Test Using Glottal Spectral Slope as the Feature

Detection Method	Vector Length	**Error Rate**: Stress Style of Test Speech **SPECTRAL SLOPE**					
		Neutral	Angry	Neutral	Loud	Neutral	Lombard
Optimal Detection	1	33.33	36.78	41.38	41.72	42.76	42.07
	5	25.45	21.82	30.91	34.55	30.91	36.36
	10	25.00	17.86	35.71	35.71	28.57	32.14
Distance Measure	5	34.55	18.18	38.89	35.19	38.89	33.33
	10	35.71	17.86	44.44	25.93	44.44	25.93

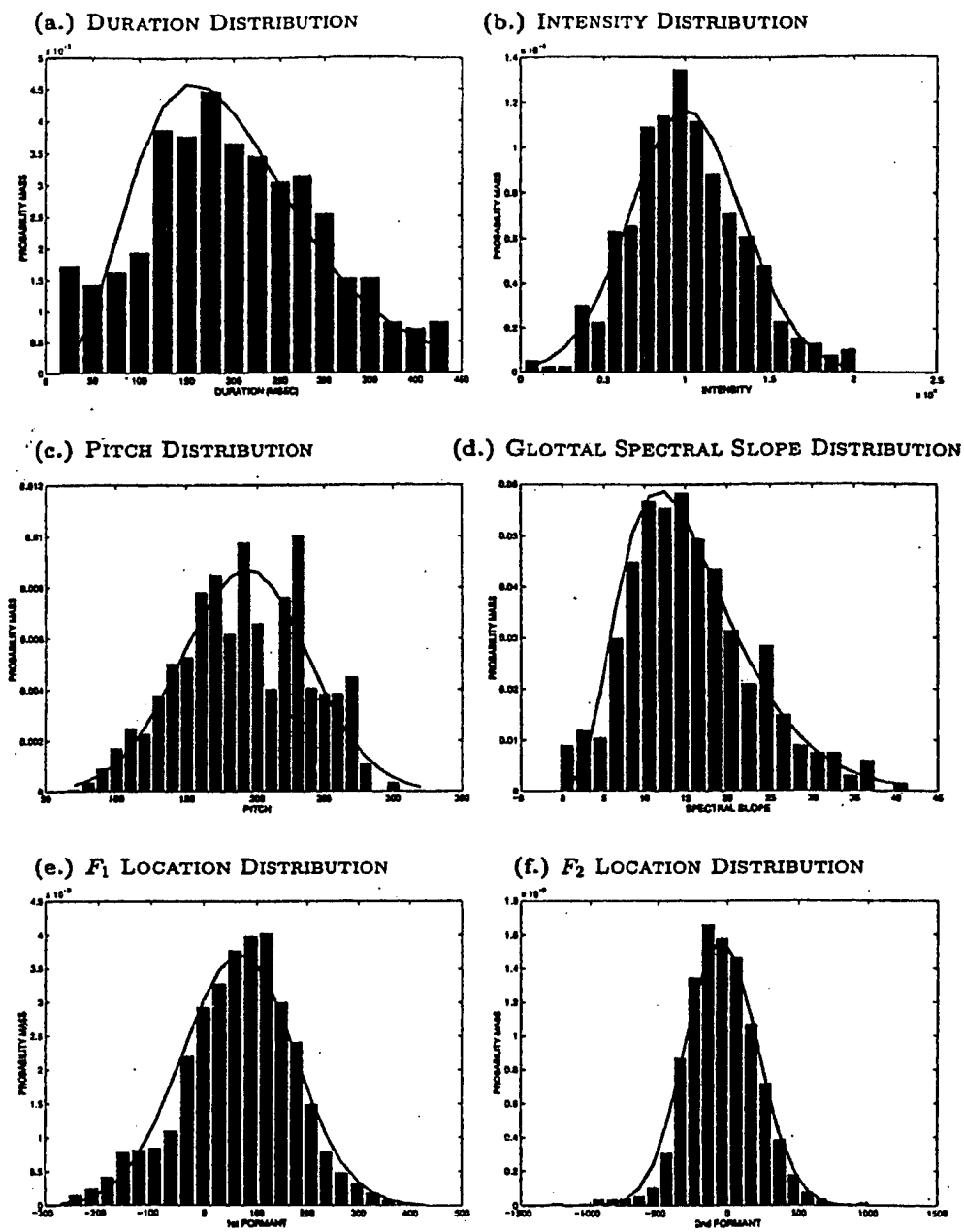
The open-set test results from the Bayesian hypothesis-testing method show that spectral slope is more suitable for detecting angry speech than for detection of loud or Lombard effect speech from neutral. In spite of this, it still does not produce a reasonable levelof accuracy for classifying angry versus neutral speech. One possible reason for this result is that a more direct glottal source estimation method might be needed, since the results presented in (Hansen, 1998b) seem to suggest that glottal spectral slope should be more successful. The distance measure approach shows a similar level of performance as that obtained using the Bayesian hypothesis-testing method.

4.2.5 Vocal Tract Spectrum

In the evaluation of vocal tract spectral structure, first and second formant location was used. Since it was of interest to reduce vowel phoneme dependent traits (i.e.,the absolute vowel formant location), formant location measurements were made with respect to the deviation from the expected average value. Therefore, using the expected average formant locations obtained from (Deller, Hansen, and Proakis, 1999; page 125), we subtract off the expected formant location knowing the particular vowel data under test (single and uppercase ARPABET labels are used for phonemes from Deller, Hansen, and Proakis, 1993; page 118). Using this conversion, formant location deviations of all vowels can be collected into a histogram and were shown to fit well to a Gaussian PDF (shown in Fig. 7e,f for first and second formants). Fig. 8e,f shows ROC curves for the Bayesian hypothesis-testing method for different vector lengths. The open-set test results are summarized in Table 7. A comparison of ROCs and distance measure performance, we conclude that first and second individual formant location are not suitable for stress detection.

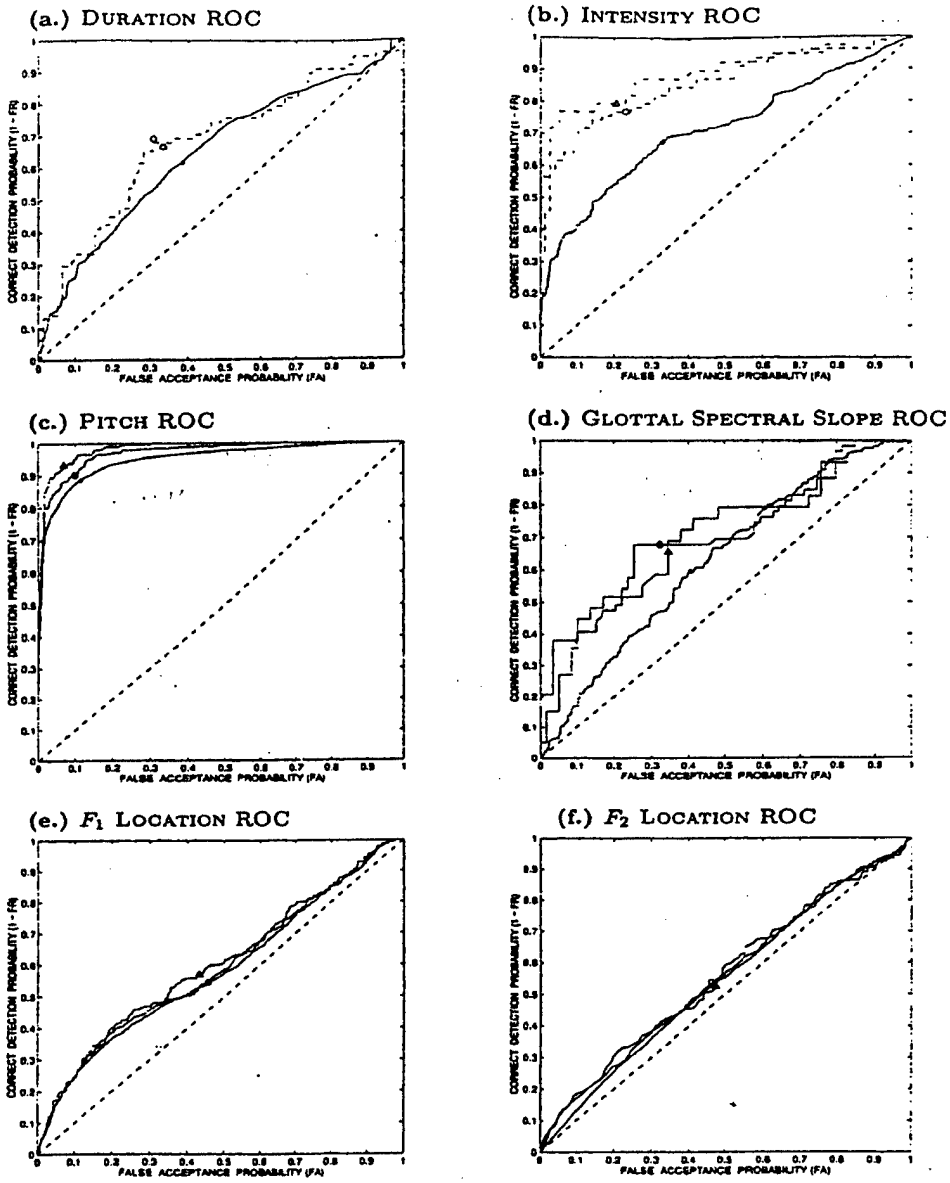
Table 7: Error Rate (percentage) of Open-set Stress Detection Test Using First and Second Formant Location as the Features

Detection Method	Formant	Vector Length	**Error Rate:** Stress Style of Test Speech FORMANT FREQUENCY					
			Neutral	Angry	Neutral	Loud	Neutral	Lombard
Optimal Detection	1st	1	42.60	41.80	46.43	45.10	46.84	46.90
		5	40.60	40.30	46.12	45.82	47.91	46.87
		10	38.79	40.91	43.03	44.24	47.58	47.88
	2nd	1	51.48	50.88	58.20	54.51	52.98	49.88
		5	53.88	49.85	58.51	56.12	54.78	50.90
		10	55.76	47.58	59.39	57.27	53.33	55.15
Distance Measure	1st	5	43.58	37.76	44.63	45.97	45.82	46.72
		10	41.82	39.39	43.64	41.82	45.45	46.06
	2nd	5	53.28	49.85	41.49	74.78	36.87	74.93
		10	54.55	49.09	40.00	74.85	38.48	76.06



Gaussian and Gamma *pdfs* used to approximate the feature distribution of vowels under loud speaking style. (a.) duration: $\Gamma(\alpha, \beta)$ with ($\alpha = 4.4402, \beta = 45.6920$);
(b.) intensity: $N(\mu, \sigma^2 | X \geq 0)$ with ($\mu = 9.99 \times 10^3, \sigma = 1.16 \times 10^7$);
(c.) pitch: $N(\mu, \sigma^2 | X \geq 0)$ with ($\mu = 192$ Hz, $\sigma^2 = 2094$);
(d.) glottal spectral slope: $\Gamma(\alpha, \beta)$ with ($\alpha = 4.2329$, $\beta = 3.6612$);
(e.) first formant location: $N(\mu, \sigma^2)$ with ($\mu = 73.28, \sigma = 1.32 \times 10^3$);
(f.) second formant location: $N(\mu, \sigma^2)$ with ($\mu = -39.90, \sigma = 6.89 \times 10^4$).

Figure 7: Gaussian and Gamma *pdfs*



ROC detection curves for "loud" versus neutral speech (vowels) using input vector lengths of (1,5,10) represented as (solid line *:, dashed line ∘, dotted line △) for:
(a.) duration: EER(*) = 38.32%; EER(∘) = 30.77%; EER(△) = 33.33%
(b.) intensity: EER(*) = 32.74%; EER(∘) = 23.08%; EER(△) = 20.51%
(c.) pitch: EER(*) = 11.47%; EER(∘) = 9.86%; EER(△) = 6.80%
(d.) glottal spectral slope: EER(*) = 40.51%; EER(∘) = 32.22%; EER(△) = 34.48%
(e.) first formant location: EER(*) = 45.67%; EER(∘) = 45.51%; EER(△) = 43.07%
(f.) second formant location: EER(*) = 46.94%; EER(∘) = 46.32%; EER(△) = 47.49%

Figure 8: ROC detection curves

4.2.6 Discussion of Linear Speech Features

Based on Tables 3, 4, 5, 6, and 7, the following observations can be made: (1) that pitch is the best feature for stress classification among the five features considered, (2) error rates generally decrease as feature vector length increases, (3) performance differences exist between different stress styles, and (4) mean vowel formant locations are not suitable for stress classification. The results in this section have therefore established stress classification performance using linear speech production based features with two types of optimum detection methods.

4.3 Stress Classification Using Nonlinear Speech Features

In this section, recently proposed approaches to stress classification that employ Teager Energy Operator (TEO) based processing are considered. Three were proposed in the study by Zhou, Hansen, and Kaiser (1998a), and a fourth was discussed in Zhou, Hansen, and Kaiser (1998b). Here, we briefly consider the basic principles of the TEO, and one nonlinear feature for stress classification (TEO-CB-Auto-Env). This is followed by evaluations using stressed speech data from SUSAS for classification. Finally, we consider a comparison of three features for stress assessment in speech using actually emergency data provided by NATO IST/TG-01.

4.3.1 Teager Energy Operator

According to studies by Teager (1980, 1983), the assumption that airflow propagates as a plane wave in the vocal tract may not hold, since the flow is actually separated and concomitant vortices are distributed throughout the vocal tract. Based on the theory of the oscillation pattern of a simple spring--mass system, Teager developed an energy operator to measure the energy for simple sinusoids which has been suggested as being a useful element for speech. The simple and elegant form of the operator was introduced by Kaiser (1990, 1993) as,

$$\begin{aligned}\psi_c[x(t)] &= \left(\frac{d}{dt}x(t)\right)^2 - x(t)\left(\frac{d^2}{dt^2}x(t)\right) \\ &= [\dot{x}(t)]^2 - x(t)\ddot{x}(t),\end{aligned} \tag{30}$$

where $\psi[\cdot]$ is the Teager Energy Operator (TEO), and $\chi(t)$ is a single-frequency component of the continuous speech signal. Kaiser (1990, 1993) derived the operator for discrete-time signals from its continuous form $\psi_c[\chi(t)]$, as,

$$\psi[\chi(n)] = \chi^2(n) - \chi(n+1)\chi(n-1), \tag{31}$$

where $\chi(n)$ is the sampled speech signal.

The TEO is typically applied to a bandpass filtered speech signal, since its intent is to reflect the energy of the nonlinear energy flow within the vocal tract for a single resonant frequency. Under this condition, the resulting TEO profile can be used to decompose a speech signal into its AM and FM components within a certain frequency band via,

$$f(n) \approx \frac{1}{2\pi T}\arccos\left(1 - \frac{\psi[y(n)] + \psi[y(n+1)]}{4\psi[\chi(n)]}\right), \tag{32}$$

$$|a(n)| \approx \sqrt{\frac{\psi[\chi(n)]}{\left[1 - \left(1 - \frac{\psi[y(n)] + \psi[y(n+1)]}{4\psi[\chi(n)]}\right)^2\right]}}, \tag{33}$$

where $y(n) = \chi(n) - \chi(n-1), \psi[\cdot]$ is the TEO operator as shown in Eq. 31, $f(n)$ is the FM component at sample n, and $\alpha(n)$ is the AM component at sample n. On the basis of this work, Maragos, Kaiser, and Quatieri (1993a,b) proposed a nonlinear model which represents the speech signal $s(t)$ as,

$$s(t) = \sum_{m=1}^{M} \tau_m(t), \tag{34}$$

where

$$\tau_m(t) = \alpha_m(t)\cos\left(2\pi\left(f_{cm}t + \int_0^t q_m(\tau)d\tau\right) + \theta\right) \tag{35}$$

is a combined AM and FM structure representing a speech resonance at the mth formant with a center frequency $F_m = f_{cm}$. In this relation, $\alpha_m(t)$ is the time-varying amplitude, and $q_m(\tau)$ is the frequency modulating signal at the mth formant.

Although the TEO is formulated for single-frequency signals or signals with a single resonant frequency, previous studies have shown that the TEO energy of a multi-frequency signal is not only different from that of single-frequency signal but also reflects interactions between different frequency components (Zhou, Hansen, and Kaiser 1998a,b). This characteristic extends the use of TEO to speech signals filtered with wide bandwidth band-pass filters (BPF). In the next section, we consider one TEO based features for stress classification.

4.3.2 TEO-CB-Auto-Env: Critical Band Based TEO Autocorrelation Envelope

Empirically, the human auditory system is assumed to be a filtering process which partitions the entire audible frequency range into many critical bands (Yost, 1994). Based on this assumption, a nonlinear feature is proposed that employs a critical band based filterbank to filter the speech signal followed by TEO processing (see Fig. 9) Each filter in the filterbank is a Gabor bandpass filter, with the effective RMS bandwidth being the corresponding critical band. This feature is an extension to previous TEO based features which have been proposed (Zhou, Hansen, and Kaiser 1998a), and preliminary classification results have also been reported (Zhou, Hansen, and Kaiser 1998b). Here we consider a comparison with other features for classification, and extend the basic ideas for the problem of stress assessment.

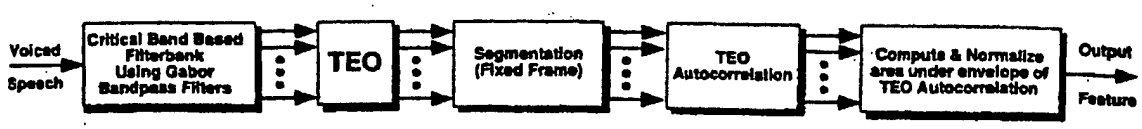


Figure 9: TEO-CB-Auto-Env Feature Extraction

To extract the TEO-CB-Auto-Env feature, each TEO profile of a Gabor BPF output is segmented into 200-sample (25 msec) frames with 100-sample (12.5 msec) overlap between adjacent frames. Next, M normalized TEO autocorrelation envelope area parameters are extracted for each time frame (i.e., one for each critical band), where M is the total number of critical bands. This is the TEO-CB-Auto-Env feature vector per frame. Fig. 9 shows the entire feature

extraction procedure. Since each critical band possesses a much narrower bandwidth than the 1 kHz bandwidth used for BPFs in the TEO-Auto-Env feature (discussed in Zhou, Hansen, and Kaiser 1998a), post Gabor bandpass filtering centered at median $F0$ is not needed in TEO-CB-Auto-Env extraction. This makes the new feature independent of the accuracy of median $F0$ estimation.

In practice, all TEO profiles are segmented into many frames and all autocorrelation functions are normalized. As a result, the constant autocorrelation function is represented as a decaying straight line from $(0,1)$ to $(N,0)$, where N is the frame length. Those variations caused by harmonic distribution as well as by modulations from stress are expected to be reflected by the change in the TEO autocorrelation envelopes.

4.4 TEO Based Stress Detection Evaluations

Evaluations were also conducted using the SUSAS, *Speech Under Simulated and Actual Stress* database (see Hansen, 1998a for a discussion). In experiments discussed here, angry, loud and Lombard effect styles were used from SUSAS for simulated stress (speakers were requested to speak in that style; 85 dB SPL pink noise played through headphones was used to simulate the Lombard effect). Data for SUSAS actual stress was selected from the subject motion-fear domain. In the actual domain, a series of controlled speech data collection experiments were performed with speakers riding an amusement park roller coaster.

Since the TEO is more applicable for the voiced sound than for the unvoiced sound, only voiced sections of all word utterances were used for the evaluation. A baseline 5-state HMM-based stress classifier with continuous Gaussian mixture distributions was employed for the evaluations. For the purposes of comparison, a frame based pitch and MFCC features (Davis and Mermelstein, 1980) were used.

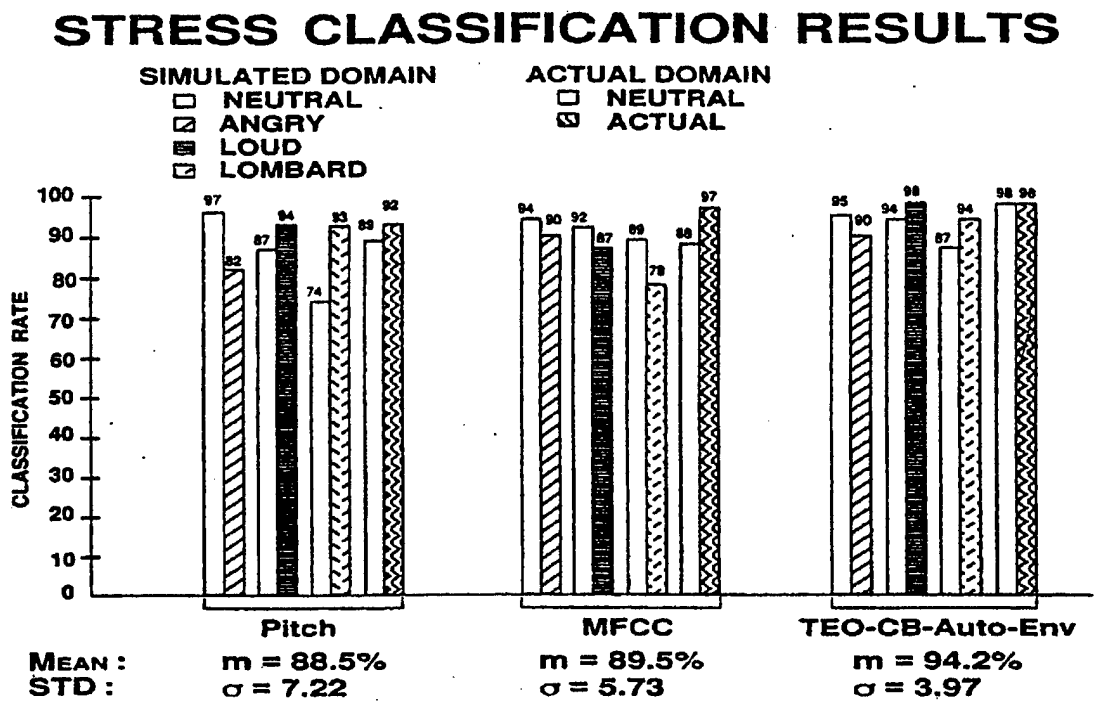


Figure 10: Pairwise Stress Classification Results (Mean and standard deviation of overall neutral/stress classification rates are shown; Different speaker groups were used for simulated and actual stress conditions)

The evaluation results are shown in Fig. 10. In general, the TEO based feature was effective in classifying stressed speech from neutral for both simulated and actual stress situations. We should expect that the performance for the neutral versus actual stress domain to be better than simulated domain (angry, loud, Lombard effect), since the speakers clearly demonstrated extreme levels of stress for this data. The TEO-CB-Auto-Env feature with its fine frequency band partitions, provides the most effective and consistent level of stress classification performance compared with MFCC and pitch information.

The evaluations in this section have shown that the proposed nonlinear based TEO-CB-Auto-Env feature is effective in the classification of speech under stress in both simulated and actual stress settings. This assumes that the goal is to detect the presence of stress. In some voice

communication settings, it is also necessary to assess the level of stress in a speaker's voice. The next section considers both linear and nonlinear based features for the task of stress assessment using actual emergency military voice communications between aircraft pilots from the SUSC-0 stress database.

5 Stress Assessment

In many commercial, law enforcement, and military applications, it is necessary to assess whether or not, as well as the degree to which, a speaker is under stress. To evaluate the techniques discussed and their ability to detect real stress, the SUSC-O database containing speech of pilots under stress was processed (in a later section, we present an equivalent evaluation of speech data from the Mt. Carmel law enforcement encounter). The SUSC-0 database is from NATO IST-TG01, which consists of actual aircraft pilot communications under emergency situations[8]. Specifically, the *Mayday2* domain in SUSC-0 was used, which contains speech data between a pilot and controller collected from the initial ground aircraft system check, through preliminary discovery of engine emergency, until safe resolution of the emergency. The different stress degrees experienced by the pilot are reflected by his speech in *Mayday2*. Twelve (12) sentences from the SUSC-0 database were extracted to represent different speaking styles for the assessment evaluation. Table 11 shows the 12 sentences from SUSC-0, where No. 1 represents ground systems check; in sentences 2-7 the pilot understands there is a problem and is working through a series a checks to determine the cause and to attempt to remedy the cause; sentence 7-11 the pilot realizes now that he is in an extreme emergency and stands a real possibility of not being able to land his aircraft; finally in No. 12 he has landed his aircraft and expresses relief.

A baseline HMM-based stress assessor with continuous Gaussian mixture distributions was used for the evaluation. Two reference HMM models, one representing neutral speech and the other representing stressed speech, were trained. All voiced segments of the word "help" under neutral conditions in SUSAS database were used to train the neutral HMM reference model. For the stressed HMM reference model, two different data sets were trained, one from a combination of simulated angry, loud, and Lombard stress conditions, and one from that actual stress roller coaster and free fall ride data, respectively. If a speech feature can assess the degree of stress regardless of text, the log likelihood ratio of the unknown speech generated by the stressed

[8] Sample audio files for stressed speech databases used in this study, SUSAS and SUSC-0, are available from the NATO IST/TG-01 Web page on *Speech Under Stress*: http://cslu.colorado.edu/rspl/STRESS/info.html

HMM model versus the neutral HMM model should be able to indicate whether it is more likely under stress or neutral. Since the TEO-based autocorrelation envelope feature (TEO-CB-Auto-Env), MFCCs, and frame-based pitch information were shown to be very effective for stress classification, they were used to assess the stress for SUSC-0 data. Since both the TEO-based feature and pitch information are only useful for voiced speech, the assessment is based on the extracted voiced portions from each utterance. To consider the variations within each utterance, 4 voiced portions per utterance (shown in Table 11) are extracted for the assessment. Note that the neutral and stress HMM classification models were trained from the /eh/ phoneme in *help*, and that almost all tested voiced sections consisted of different phonemes.

Table 8: Sentences from SUSC-0 used for Stress Assessment Evaluation. Note that bold uppercase characters represent voiced sections which were used for overall stress assessment of that sentence.

Sentences used for Evaluation from Mayday2 Domain of SUSC-0		
No.	Sentence	Extracted Phonemes
1	Avionics **lIGHt** hydr**AU**lic oil pressure **lIGHt** engine indications **ARE** ...	/ay/ /ao/ /ay/ /aa/
2	**AND** you'er **gONNA** declare an em**ERg**ency or am **I**	/ae/ /aa-n-ax/ /er/ /ay/
3	... checklist **OIl** pressure malfunction **G** one-hundred ... cruise altitude st**ORe** jett ... throttle minimize m**Ov**ement ...	/oy/ /iy/ /-/ /-/ /-/ /-/ /ao/ /uw/
4	Roger that **OIl** indic**A**tor is n**OW** z**ERO**	/oy/ /ey/ /aw/ /ih-r-ow/
5	... **ALRIGHt** newt ... engine fault **lIGHt** still lit ... hydr**AU**lics are ... total p**OUN**ds six ...	/ao-l-r-ay/ /ay/ /ao/ /-/ /-/ /-/ /-/ /aw-n/
6	And I'm going there and I'm there I'm desc**EN**ding down to ten gr**AN**d right I'm n**O**t picking up a t**A**can lock	/eh/ /ae-n/ /-/ /-/ /-/ /-/ /aa/ /ey/
7	No I'**M** doing **ALRIGHt** now and the r**A**dial is wh**A**t	/eh-m/ /ao-l-r-ay/ /ey/ /ax/
8	Ok**AY** give me imm**E**diate vectors this is an em**ERg**ency I'm gengine **OUt**	/ey/ /iy/ /er/ /aw/
9	g**I**ve me h**EA**dings I n**EE**d headings n**OW**	/ih/ /eh/ /iy/ /aw/
10	Put the c**A**ble d**OWN** p**U**t the c**A**ble down	/ey/ /aw-n/ /uh/ /ey/
11	I'm h**O**t I n**EE**d the c**AbL**e ...	/aa/ /iy/ /ey/ /ax-1/
12	m**A**n I th**OUGH**t I w**A**s g**O**ne	/ae/ /ao/ /aa/ /ao/

The assessment results are shown in Fig. 14. Here, a single score is obtained by finding an average output score across the four extracted voiced sections per sentence. Generally speaking, the recordings begin in a neutral relaxed setting (sentences 1-2), then move into concern while pilot begins to determine the cause of the problem (sentences 3-7). Finally, the pilot determines that the emergency is serious and must land the aircraft without power (sentences 8-11). Sentence number 12 indicates his relief after a safe landing.

Both figures ((a) and (b) in Fig. 14) show that the general assessment score trend is similar regardless of which anchor stress HMM reference model is used (note that a negative likelihood

difference score means that the `neutral' HMM model is more likely, and that a positive score means the `stressed' HMM model is more likely). The results do show the stress HMM reference model trained from actual SUSAS stressed speech has larger fluctuations among assessment scores. This may be because that model represents an extreme case of stress. It is noted that SUSC-0 recordings can at times have high levels of background noise, so it is possible that stress assessment could be affected by this distortion[9]. The stress level profile versus increasing sentence location showed limited variation for MFCC features. This occurs, because while there are significant changes in spectral structure on a per phoneme basis as demonstrated in (Hansen, 1998b), the differences in phoneme content for the voiced sections analyzed are more dissimilar to either neutral or stressed MFCC trained HMM model (this explains why the difference in log-likelihood scores are close to zero, since both models give similar scores). For pitch (fundamental frequency) versus time, we see that the neutral model is selected for sentence counts 1-7, with a sharp change towards the stressed HMM model for 8-12. We note that for sentence example 9, there were irregular pitch values resulting from the pitch estimation scheme which were not corrected (i.e., we wanted to compare performance of features without user intervention). Finally, the TEO-CB-Auto-Env feature produced more meaningful scores for the case of a neutral versus Actual stress trained reference HMM model as opposed to a simulated stress trained reference HMM. Again, the neutral model received very high scores (large negative likelihood difference score) for sentence entries 1-7. Sentence entries 8-12 produced scores which were more associated with the stressed model in both test cases. Since the neutral reference HMM model was the same in both test cases, the difference in scores reflect differences in the stressed reference model. The results here demonstrate that the proposed feature can be used for the purposes of stress assessment, though it is suggested that the stressed speech reference model should be trained on data which reflects the desired type of stress to be assessed. Also, future studies could consider the influence of other distortions for assessment, including channel/microphone differences and acoustic background interference.

[9] In this study, we choose not to perform speech enhancement due to the potential of introducing spectral based processing artifacts (Hansen 1999).

A second evaluation for stress assessment will be presented in Section 7, which specifically considers the law enforcement voice recordings from the shoot-out at Mount Carmel.

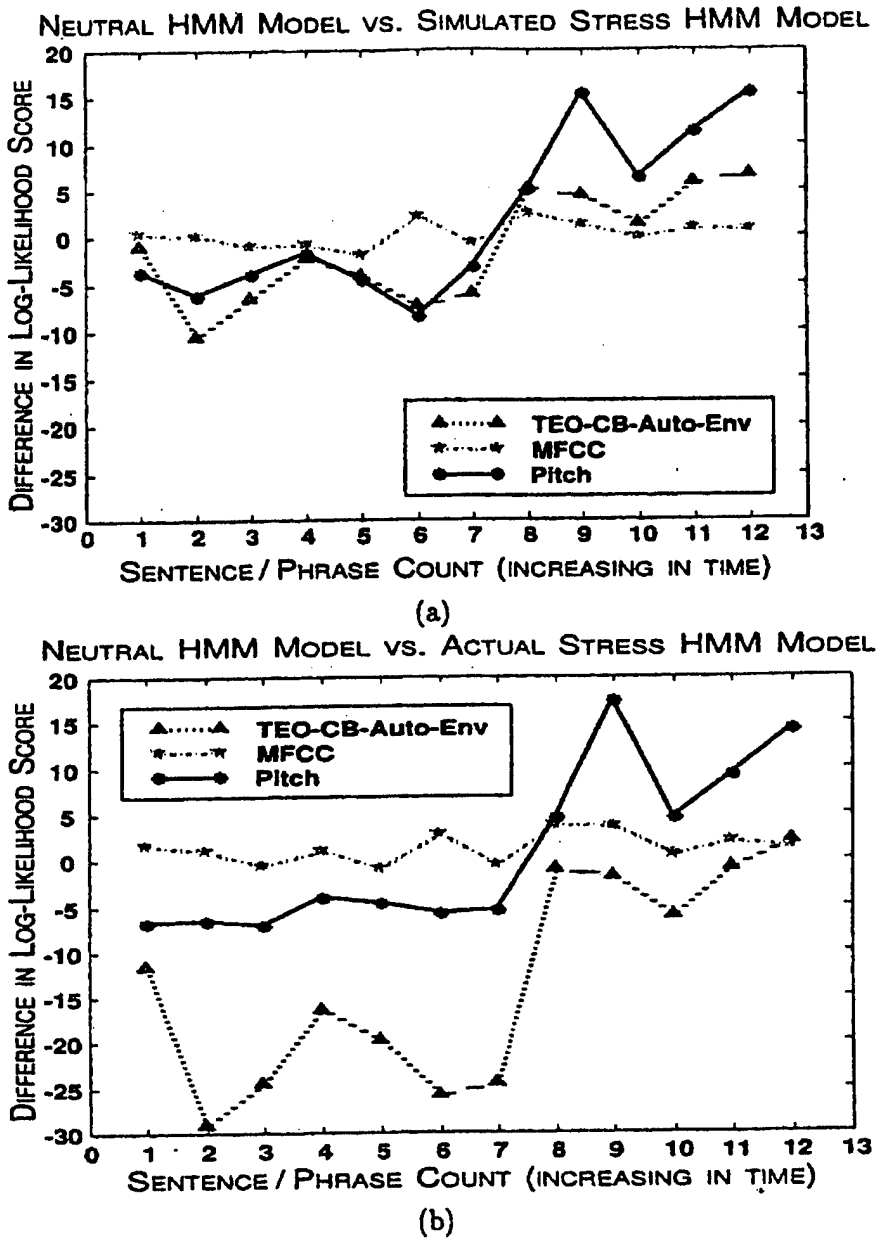


Figure 11: Assessment results for pilot's speech from Mayday2 domain of SUSC-0 database (Log likelihood ratio is shown along Y-axis while sentence number is shown along X-axis): (a) Neutral vs Simulated stress (Loud, Angry and Lombard) HMM reference models; (b) Neutral vs Actual stress HMM reference models

6 Conventional/Commercial Voice Stress Analyzer Features

In this section, we consider an evaluation of several traditional features which have been used in the development of commercial voice stress analyzers. The results here are presented in the form of a series of experiments. The three features considered include: (i) normalized pitch frequency, (ii) periodicity, and (iii) pitch jitter. The evaluations were conducted using three stressed speaking styles extracted from the SUSAS speech database. The stressed speech conditions include: Angry, Loud, and Lombard effect.

6.1 Features: Normalized Pitch, Periodicity, Jitter

The scaled pitch measure is computed using the autocorrelation method. For the *i*th frame of windowed speech, $s_i(n)$, the maximum valued autocorrelation lag, $m_{\max}(i)$ is computed using the function,

$$m_{\max}(i) = \operatorname{argmax}\left\{R_i(m) = \frac{1}{N-m}\sum_{\ell=0}^{N-m-1} s_i(\ell)s_i(\ell+m)\right\}. \tag{36}$$

The pitch frequency of the signal is obtained by dividing the sample rate, F_{sample}, by the maximum valued autocorrelation lag,

$$F_o(i) = \frac{F_{sample}}{m_{\max}}. \tag{37}$$

Finally, scaled pitch is obtained by first applying the constraint that $\left\{80Hz \le F_o(i) \le F_0^{\max}\right\}$ and dividing by a maximum allowable pitch frequency $\left(F_0^{\max} = 400Hz\right)$,

$$\tilde{F}_o(i) = \frac{F_o(i)}{F_0^{\max}}. \tag{38}$$

Scaled pitch values range from 0 to 1 with values near 1 typically observed for speech under extreme stress.

Periodicity represents the degree of voicing state of the speech waveform. It is simply computed as the ratio of the energy of the $m_{\max}$ autocorrelation lag:

$$P(i) = \frac{R_i(m_{\max})}{R_i(0)} \tag{39}$$

Jitter is related to the frame-to-frame variation in pitch period and essentially measures small fluctuations in glottal cycle lengths. Let $V(i)$ represent the absolute difference between the pitch period at frame i and frame i-1:

$$V(i) = |P(i) - P(i-1)| \tag{40}$$

Jitter, $J(i)$ is computed as follows

$$J(i) = \frac{\frac{1}{2}[V(n) - V(n+1)]}{\frac{1}{3}[P(i-1) + P(i) + P(i+1)]} \tag{41}$$

6.2 CVSA: Computer Voice Stress Analyzer

The operation of the computer voice stress analyzer (CVSA) is based on the notion that muscles and limbs of the human body exhibit a natural tremor rate ranging between 8 to 12 Hz. There are several underlying assumptions made about speech production which leads to the formulation of the device. First, since vocal chords are primarily muscular tissue, it is assumed that the voice fundamental frequencies are modulated by an 8 to 12 Hz "microtremor". Second, increased levels of arousal or stress contribute to additional tension in the vocal chords. This results in a reduction of the natural tremor amplitude. Finally, it is assumed that "microtremors" are not audible to the listener, but measurable using computer aided algorithms.

Various devices have been constructed to measure microtremors in the human voice. The analog device known as the Psychological Stress Evaluator (PSE) was studied by VanDercar, et. al (1980). The general operation of the device consists of four basic modes. Each operation mode (known as Mode 1 to Mode 4) controls the degree to which the signal is filtered. The filtering in all four modes was accomplished using a combination resistor and capacitor circuit to produce

varying degrees of low-pass response. In Mode 3, for example, the PSE circuitry consists of a 4.6 μF capacitor in parallel with a 30KΩ resistor.

After reviewing the literature for the CVSA as described in (Cestaro, 1995), we implemented a Matlab version of the CVSA. The Matlab software is very simple in that it implements the digital filter described by Mode 3 operation of the PSE device. The software assumes input speech sampled at 8 kHz and outputs a time-domain waveform shape analogous to the pen-drawings illustrated in (VanDercar, et. al, 1980).

During processing, the speech signal is first passed through an 8 times oversampling to simulate the one-eighth tape play speed of Mode 3. After oversampling, the waveform is passed through a low-pass digital filter with frequency response derived from the resistor/capacitor description of the analog device. The Matlab code listing is shown below:

```
function z = cvsa(x)

  samp = 8000;
  pass = 12;
  stop = 15;

  x = resample(x,8,1);
  x(find(x<=0)) = zeros(length(find(x<=0)),1);
  y  = x;
  [n,Wn,beta,typ] = kaiserord([pass stop],[1 0], [0.01 0.1], 8000 );
  b = fir1(n, Wn, typ, kaiser(n+1,beta), 'noscale');
  z = filter(b,1,y);

  plot(z);
```

The CVSA output is analyzed visually. Four aspects of the output waveform are assumed to contribute to reveal the degree of vocal stress. These include: amplitude, leading edge, cyclic rate change, and "blocking". A description of each term and it's visual manifestation can be found in (VanDercar, et al., 1980). The most important indicator of stress or deception in speech is thought to be "blocking". Blocking occurs when straight parallel lines are seen in the output to form an envelope over the CVSA signal. Evaluation of the implemented CVSA scheme is presented in Section 7.

6.3 Evaluations: Normalized Pitch, Periodicity, Jitter

The speech data consisted of simulated stress from the SUSAS speech database. Specifically, 56 isolated words from each of 9 speakers were used to estimate GMM (Gaussian Mixture hidden Markov Model) based models for each stressed condition. The remaining 14 words were used for open test evaluation. Due to limited data, a round-robin train/test paradigm was used. During processing, each word token was first processed using an automatic end-point detection algorithm. Next, the (3) features were extracted every 10 msec from 30 msec windowed portions of data.

The evaluation consisted of a pair-wise stress classification task. Data submitted for test was assumed to be either neutral data or one of three stressed speaking styles. The classifier must therefore decide if the data is either neutral or stressed.

The evaluation consisted of submitting the test set data (different from training data) to each GMM (normal, angry, loud, Lombard). The output scores for each frame were used to compute a frame-based log likelihood ratio. The average of the frame-based measures were computed over a single isolated word and the output compared to a decision threshold. Values greater than the threshold are considered to be from normal speaking conditions while values less than the threshold constituted stressed speaking style. The results summarized below are presented in the form of a series of experiments which serve to determine if the GMM classifier structure, or the input speech data type, influence stress classification performance.

Experiment 1: In order to determine the influence of the number of mixtures in the GMM classifier, we ran an experiment with three different mixture sizes. All three features (i.e., normalized pitch frequency, periodicity, and pitch jitter) were used as a per frame vector. The results are shown in Table 9. In general, as the number of Gaussian mixtures is increased, the ability of the classifier to more closely represent the changing feature structure should increase. As the results in Table 9la show, there is only a slight increase in performance as the number of mixtures increase. Since excitation features change more significantly for angry and loud speech, we would expect their performance to be much better than for Lombard speech. While

this is true, the difference is not as large as one might expect if we simply considered mean pitch changes.

Table 9: Experiments using (i) normalized pitch frequency, (ii) periodicity, and (iii) pitch jitter as a three feature set for a GMM (Gaussian mixture model) stress classifier. Evaluations using 3 different size sets of mixture weights, and adding first and second order feature derivatives.

Normal vs.	Pairwise GMM-Based Stress Classification Results		
(1a)	32 mixtures	64 mixtures	128 mixtures
Angry	72.1%	70.9%	71.4%
Loud	69.2%	72.2%	75.8%
Lombard	62.7%	59.6%	64.4%
(2a)	32 mixtures	64 mixtures	128 mixtures
Angry	75.0%	75.0%	73.9%
Loud	77.6%	71.7%	71.2%
Lombard	63.7%	63.3%	59.6%
(2b)	32 mixtures	64 mixtures	128 mixtures
Angry	81.3%	78.7%	80.2%
Loud	82.3%	78.5%	78.0%
Lombard	66.0%	61.0%	61.0%
(3a)	32 mixtures	64 mixtures	128 mixtures
Angry	82.8%	78.8%	80.4%
Loud	86.6	79.5%	82.3%
Lombard	67.8%	68.7%	68.7%
(3b)	32 mixtures	64 mixtures	128 mixtures
Angry	83.2%	84.0%	83.2%
Loud	86.6%	85.8%	83.6%
Lombard	70.8%	76.9%	69.2%

Experiment 2: In this experiment, the conditions are the same as that for Experiment 1, with the exception that only voiced speech sections were used in the 3-feature vector per frame. To determine which frames were voiced, we extracted all framew with a periodicity measure greater than 0.30. The results in Table 9 (2a) are for the case when the pitch mean is removed, and 9 (2b) are for the case when pitch mean is not removed. In cases where pitch mean was previously shown to change significantly (i.e., loud and angry), the stress classification results were better. The results are about the same for Lombard speech.

Experiment 3: Several experiments were also performed were we augment the three excitation features with the first and second-order derivatives. Results for Table 9 (3a) are for the case for a combined 6 feature vector (3 static, 3 first-order derivatives) in the stress classification. In this scenario, stress classification performance improves for Lombard speech, but little real

improvement is observed for angry or loud. If the second-order derivatives are included (now a 9 feature vector per frame; results in Table 9 (3b), there is a measurable level of improvement. This was especially true for the 64 mixture case, and less so for the 128 mixture case. Again, including static, along with first and second order derivatives generally provides better resolving power for the classifier.

Experiment 4: Having established a baseline system, using 64 mixtures, we set out to explore several issues involved in the training process. One issue of interest is that when different classes of features are used, quite often their variances will encompass a wide range. To reduce these effects, we set a variance threshold during the training process (two experiments were performed; one with a variance floor of 0.001 instead of the standard 0.01 (Table 10 (4a)); and one with a variance floor of 0.0001 (Table 10 (4b)). Comparing results from Table 10 (4a) with Table 9 (3b) (64 mixture column), we see that reducing the variance floor increases classification performance, with good gains for loud and Lombard stress styles. However, dropping the variance threshold too low, results in a slight loss in performance.

Experiment 5: In addition to adjustments in the feature variance floor during training, the number of iterations, given the training corpus, can also effect classification performance. Too many iterations, will result in a model that is too specialized for the training set (especially true if the training token size is small). Too few iterations will produce a classifier which is to general. Again, this issue will be based on the amount and speaker set range in the available training data. In this experiment, we kept the same configuration as that for Experiment 4a, but considered increasing the number of iterations of the traditional Baum-Welch hidden Markov model training algorithm from 10 to 20. The results are summarized in Table 10 (5). Again, the additional training iterations, coupled with the adjustment in the feature variance floor, produces another slight increase in classification performance. We also tried an experiment where we used this set-up with frames which had a higher degree of voicing to see if transitional frames between voiced and unvoiced speech had much influence in the classifier performance. The results were almost the same, thus suggesting that transitional frames do not significantly impact performance for these stress conditions.

Table 10: Experiments using (i) normalized pitch frequency, (ii) periodicity, and (iii) pitch jitter as a three feature set for a GMM (Gaussian mixture model) stress classifier. Here, test cases explore differences in the minimum feature variance during training, the number of training iterations, and augmenting excitation based features with vocalt tract spectral features (MFCCs). The last experiment considers neutral versus grouped stress conditions.

	Pairwise GMM-Based Stress Classification Results		
Normal vs.	Angry	Loud	Lombard
(4a) variance floor:0.001	86.4%	90.5%	82.0%
(4b) variance floor:0.0001	83.9%	88.9%	80.5%
(5) Training: 20 iterations	87.4%	92.9%	81.0%
(6a) with MFCCs	94.6%	95.9%	87.5%
(6b) with MFCC, deltas, delta-deltas	92.6%	95.6%	86.9%
(7) Neutral vs. grouped Stress	93.1%	96.2%	87.4%

Experiment 6: In the experiments thus far, we have considered different forms of features which represent excitation characteristics. However, it has been shown that stress also effects spectral structure as reflected in the vocal tract structure. In this experiment, we augment the three excitation features with traditional spectral based MFCC parameters, which generally reflect vocal tract structure. To help reduce the effect of glottal source information on the MFCC parameters, we performed a pre-emphasis (coefficient of 0.97). A 20 set filterbank was used to obtain 8 MFCC spectral features per speech data frame. The results in Table 10 (6) showed a marked improvement for all three stress conditions. We also considered the case where first and second order derivatives were included. In order to reduce the impact of the fine spectral structure, we reduced the number of static MFCC parameters from 8 to 4, and included 4 delta-MFCC and 4-delta-delta MFCC parameters (i.e., first and second order derivatives). The delta features reflect the time rate of change of the static spectral structure. While including delta and delta-delta MFCC parameters have been shown to improve recognition of speech under stress, there was either no change or a slight loss in stress classification performance when included. Other experiments were also performed where we increased the variance of the excitation features by a scale constant, so that they would have more influence over spectral features. The results were within 0.1% of the values obtained in Table 10 (6a) and 10 (6b).

We point out here, that the use of spectral structure assumes that we have some examples of the speaker(s) in both neutral and stressed speaking conditions. Mean normalized excitation features

generally are less speaker dependent, and therefore more appropriate for use when training speaker data is obtained from different speakers in similar test conditions.

Experiment 7: In this last experiment, we consider a test condition originally proposed by Womack and Hansen (1996), where instead of a binary stress classification decision, we assume that the speech is either neutral or stressed, and determine an overall detection rate. This essentially groups the three stress conditions into one class (we use all three stressed GMM models during the test, and if any one is selected over the neutral model, the input is classified as stressed). This decision process does not record an error if an incorrect stressed model is selected (i.e., if the input token is under angry stressed condition, and the loud stressed model is selected, then the input was correctly identified as being under stress). This scenario was chosen, because in many situations speakers are not producing speech under a single style, but in fact typically display a mixture of conditions. The results, Table 10 (7), are nearly the same as those for the case when MFCCs are included.

In summary, the best Gaussian mixture model based classifier for these stress conditions are as follows: excitation features include normalized pitch, periodicity, and jitter with their first and second order derivatives, use 20 iterations of the training algorithm, reduce the feature training variance threshold to 0.001, use 64 mixtures per model, and include at least some form of vocal tract spectral structure (MFCCs) if data is available.

7 Stress Analysis: Mt. Carmel Data

In this section, we consider analysis and evaluation of the actual stressed speech from Mt.-Carmel. In Section 7.1 example feature plots are compared between the excitation features discussed in Section 6.1 for sample SUSAS and Carmel speech data. Section7.2 considers stress assessment using pitch, MFCCs, and the nonlinear TEO based feature for the Mt. Carmel data. The Carmel data represents audio recordings between individuals during a law enforcement encounter with armed extremists.

7.1 Example Excitation Feature Plots

In the previous section, we discussed a number of experiments to determine the usefulness of traditional excitation features for stress classification. Here, we use the same Gaussian mixture model classifier trained using the Maximum Likelihood approach. The evaluation here, however, is focused on a comparison of these features with CVSA for both Mt. Carmel law enforcement data and SUSAS speech under stress data. Several frames of processed speech output for (1) Normalized pitch, (2) Jitter, and (3) CVSA output from Matlab are considered. While it is difficult to make certain judgements from only a few examples of stressed and neutral speech, we use a comparison with examples from SUSAS which contained much more test data.

The first plot (Fig. 12) shows results for telephone speech collected from a 911 call made during the FBI raid on Mt. Carmel. Here, we see that the high-stress condition results in normalized pitch values near 1 throughout the beginning and end of the audio fragment. There is also an increase in the jitter output near the middle of the segment. For the CVSA output, we see that the variations in the output waveform are reduced for the case of the high-stressed speech, which we would expect for the case when microtremors are absent due to the presence of stress. This would also, however, contradict expectations of "blocking" which should be readily visible for speech under stress.

In order to compare these results with earlier evaluations, we repeated these evaluations with speech data from SUSAS. Data from the neutral word "fix" and the same word produced under actual stress (roller coaster environment) were processed. Fig. 13 presents feature profiles for

the three features (normalized pitch, jitter, CVSA). Because the classifiers considered are statistical in nature, it is difficult to visually see significant differences between the normalized pitch and jitter features. The CVSA outputs show significant differences between the neutral and stressed conditions. However, we point out that the stressed speech signal lacks the "blocking" output that is expected from the CVSA in stressed conditions. We might point out that speech data from the actual portion of SUSAS was from roller coaster rides, which potentially could include low frequency physical vibration.

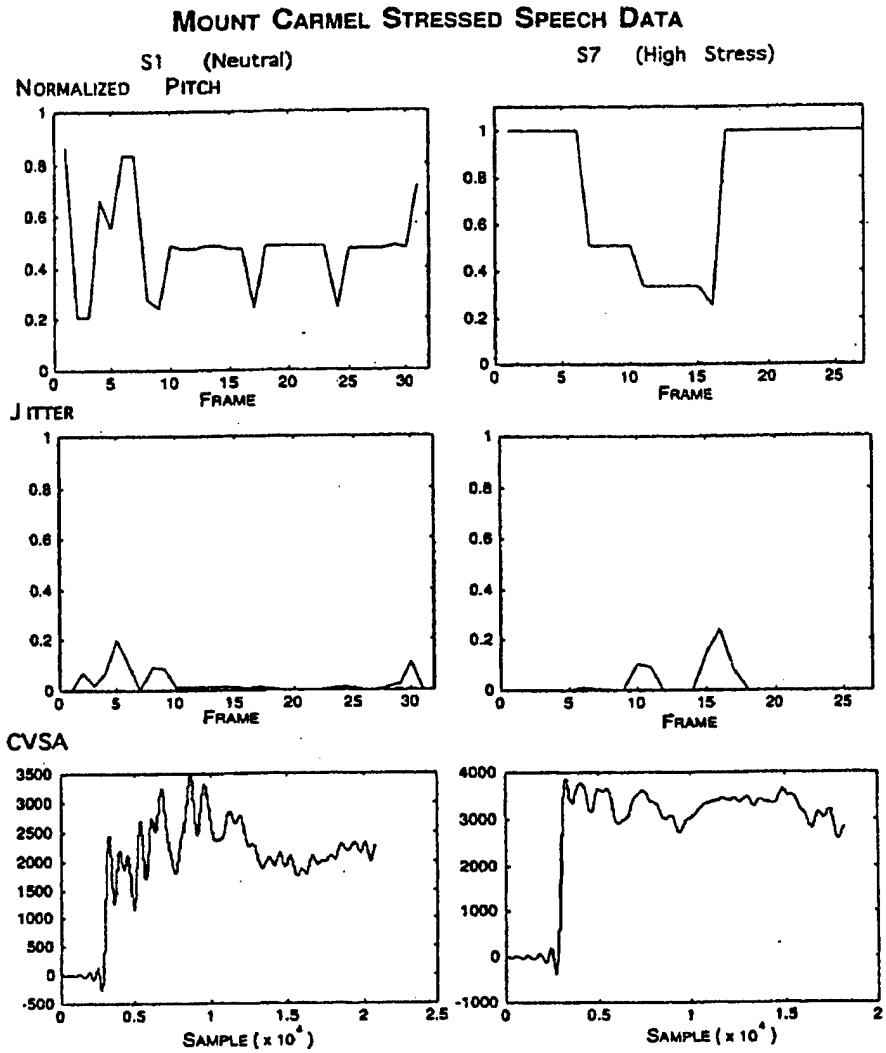


Figure 12: Feature analysis results for speech from Mt Carmel Recording. Sentence S1 and S7 were selected. Three features include (i) normalized pitch, (ii) jitter, and (iii) CVSA response.

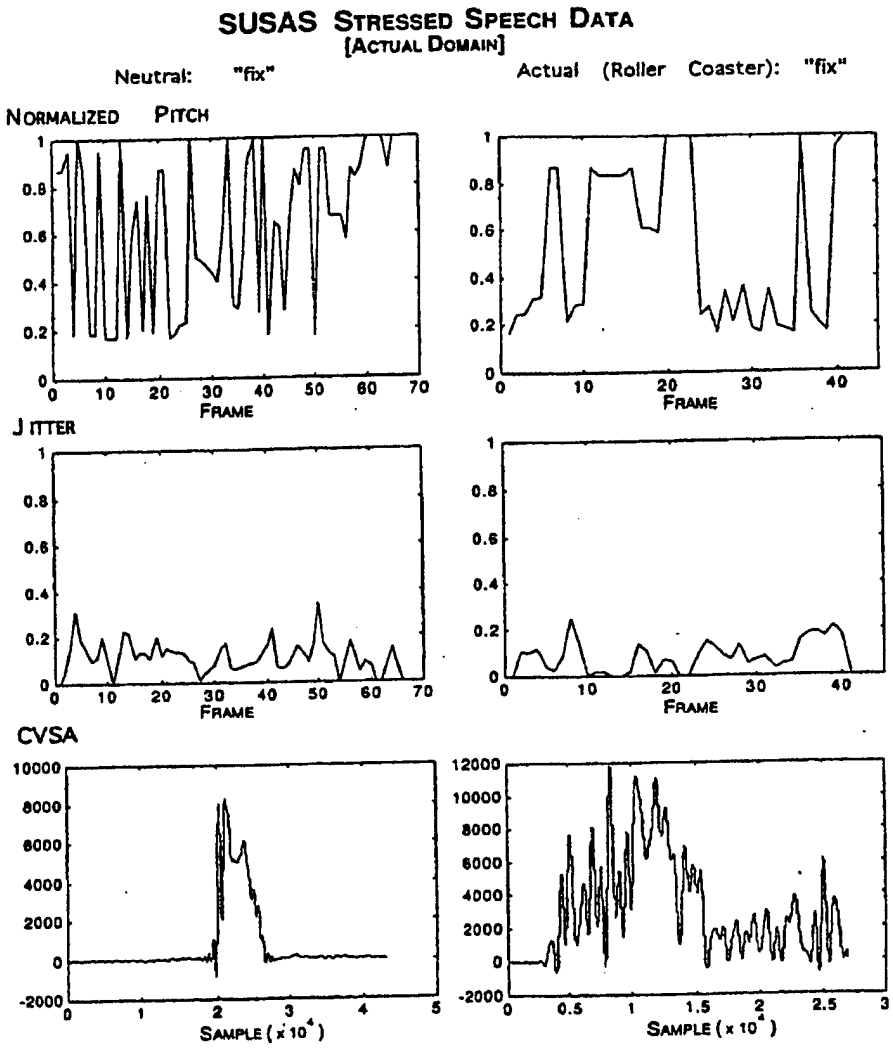


Figure 13: Feature analysis results for speech from Mt Carmel Recording. Sentence S1 and S7 were selected. Three features include (i) normalized pitch, (ii) jitter, and (iii) CVSA response.

7.2 Assessment Evaluation for Mt. Carmel Data

In this section, a stress assessment evaluation similar to that presented in Section 5 is considered, using speech data recorded during Mt. Carmel law enforcement encounter. The audio recordings obtained consisted of telephone conversations between an extremist individual (sect leader) within the compound who called 911 emergency services from the beginning of the shooting.

The speech we assessed was that of the sect leader's voice during his dialogue with the 911 service. A total of 20 sentences were segmented from the talker's speech and used for and experiment in stress assessment. In that specific situation, almost all 20 sentences were spoken under stress. However, the degree of stress varies from time to time. For example, the sect leader explained the situation in sentences 1, 9, and 10 in relatively neutral conditions; while sentences 7 and 8 were spoken during the actual shooting, with gunshots present as background noise. It is clear from these examples that the speaker was under an extreme level of stress. Similar to the experiment using SUSC-0 data, four voiced portions per utterance were extracted for assessment (text transcriptions and extracted voiced sections are summarized in Table11.

Table 11: Sentences from Mt. Carmel Recording used for Stress Assessment Evaluation. Note that bold uppercase characters represent voiced sections, which were used for overall stress assessment of that sentence.

Sentences used for Evaluation from Speech Recorded from Mt. Carmel Shooting	
No.	Sentence Text
1	There are mAn, seventy fIve men arOUNd our bUILding, ... shooting at us
2	YEAH, there are seventy five mEN around our building, they are shOOting at us at MOUNt Carmel.
3	YEAH, tell them there chILdren and women in hERE and to cALL it off
4	TELL thEM to cALL it
5	TELL thEM to pULL bAck
6	TELL thEM to pULL bAck
7	I Am UNdER fIRE
8	I have the rIght to defend mysELf. They started fIring fIRst
9	'nother chOpper with mORe people and mORe guns going on. here they cOMe..
10	That's nOt Us, thAt's thEM
11	We wAnna cEAse-fIRE, we'll tALk
12	we'll tALk when thEY stOp fIring
13	ThEY ARE,[thEY} ARE}]. (two different speakers)
14	They hAven't bEEN}, ..., thEY hAven't been (different speaker breaks in)
15	ThAt's thEM, thEY hAven't been .[??].. shooting.[???] (noise breaks in during speech)
16	They're, What do you thINk they doing ALL this fIring on us right nOW?
17	lEAst thREE (break into two portions) hIts
18	ONE (break into two) dEAd (break into two)
19	I'M tALkING [??] (another speaker breaks in at the end)
20	HOLd their fIRE}, to lEAve the property and we'll tALk

The assessment results are shown in Fig. 14. Instead of using the actual score difference for the y-axis as we did in Fig. 11, we used normalized HMM score difference. In essence, the HMM score differences are normalized for each feature, respectively, based on the corresponding range. This was performed because the range of HMM score differences for pitch was so large

that the change in score difference for the TEO-CB-Auto-Env feature and MFCC feature could not be observed clearly when all three were plotted on the same figure. As we can see, the general assessment score trend is independent upon which anchor stress HMM reference model is used (i.e., one trained using simulated stress data from SUSAS or actual stressed speech from SUSAS). Sentences assessed in this experiment have different levels and types of background noise, such as gunshots, etc. So the prospect exists that assessment results could be affected by background noise. Upon a careful listener evaluation of all 20 sentences, we found that pitch and the TEO-CB-Auto-Env feature reflected similar information regarding the degree of perceived speaker stress; while the MFCC feature was very inconsistent. We also note that the accuracy of stress assessment could be influenced by the type of recording condition. In some cases here, the speech sounds `hollow' as if the microphone recording conditions changed (there are examples where the speaker is actually yelling and cases where his mouth could be some distance from the microphone). There are also many examples where the voiced portions are very short. In spite of these observations, it appears that relative to the first sentence, there is some degree of consistency for sentences which are more relaxed and those which are under higher degrees of stress.

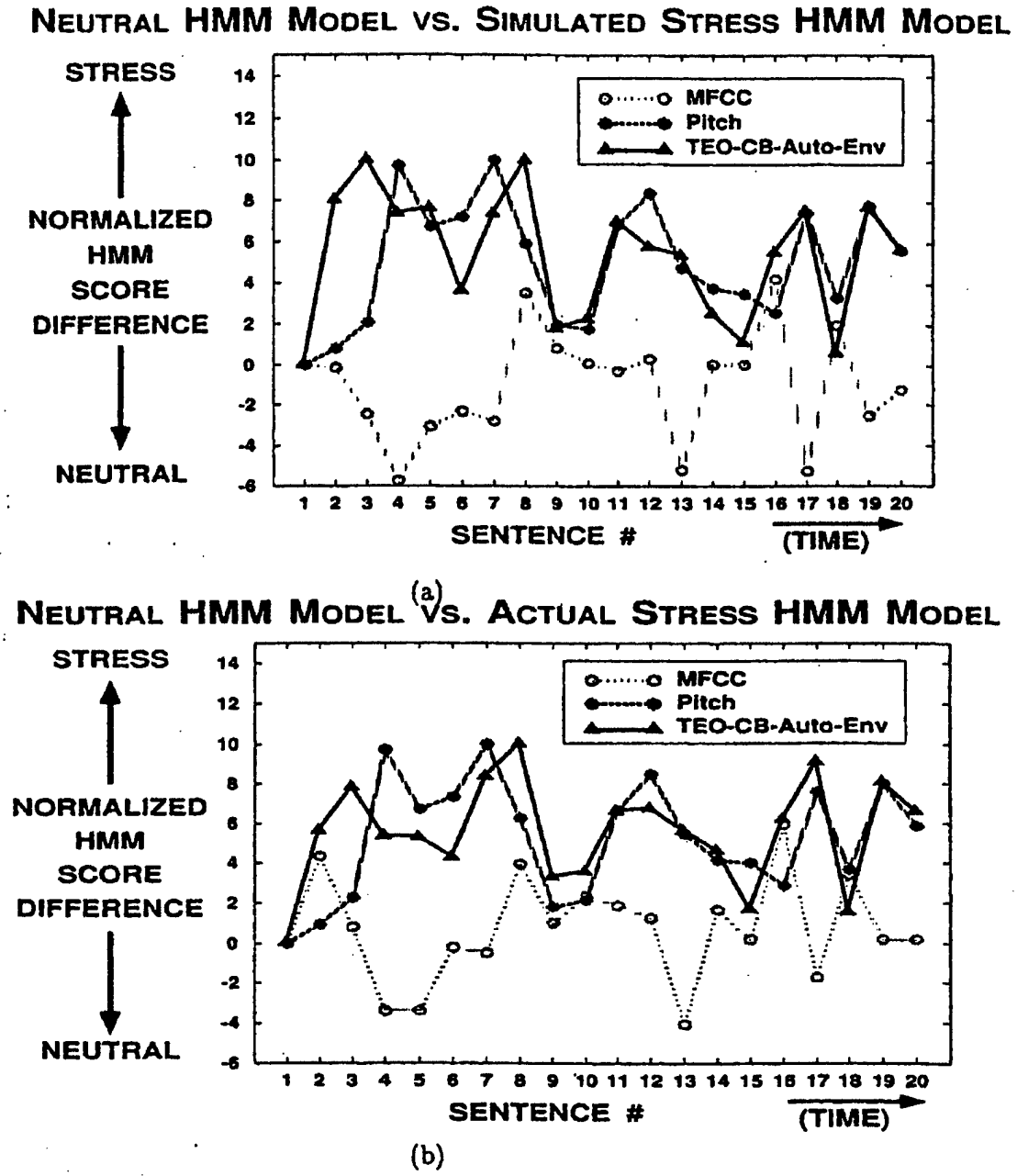


Figure 14: Assessment results for speech from Mt Carmel Recording (Log likelihood ratio is shown along Y-axis while sentence number is shown along X-axis): (a) Neutral vs Simulated stress (Loud, Angry and Lombard) HMM reference models; (b) Neutral vs Actual stress HMM reference models}

8 CONCLUSIONS: Issues for Stress Assessment and Classification

The issue of stress classification is a problem which is becoming increasingly important for law enforcement and military in the field. Past methods for voice stress analysis have focused on what is believed to be microtremors in the muscles for voice production. While there is evidence which suggests that muscle control within the speech production system could, and most likely are, influenced by the presence of stress experienced by the speaker; there is still uncertainty to what degree and how consistent this change in speech muscle control could actually manifest itself into the form of "microtremors" during the speech production process. Clearly extensive research in the medical field has considered neurological based factors that effect human speech production (for example, the work done for Parkinson's speech (L. Ramig, in Kent 1992).

In this report, we have considered previous studies on speech under stress, results from our own evaluations, experiments using features derived from commercial voice stress analyzers, and novel nonlinear based features recently formulated in the literature. All of these findings suggest that when a speaker is under stress, their voice characteristics change. Changes in pitch, glottal source factors, duration, intensity, and spectral structure from the vocal tract are all influenced in different ways by the presence of speaker stress. Our results also suggest that the features by which commercial voice stress analyzers are based upon, can at times reflect changes in the speech production system which occur when a speaker is under stress. However, as is the case with speaker control of pitch, a variety of factors could influence the presence or absence of the microtremors, which are claimed to exist in our muscle control during speech production. It is clearly unlikely that a single measure such as that based on the CVSA, could be universally successful in assessing stress (such as that which might be experienced during the act of deception). However, it is not inconceivable that under extreme levels of stress, that muscle control throughout the speaker will be affected, including muscles associated with speech production. The level and degree to which this change in muscle control imparts less/more fluctuations in the speech signal cannot be conclusively determined, since even if these tremors exist, their influence will most certainly be speaker dependent. A similar argument has been

made in the medical community over non-invasive voice analysis for screening of subjects with vocal fold cancer (Hansen, Gavidia-Ceballos, and Kaiser, 1998).

Many commercial voice stress analyzers are presently on the market. Some of these include:

- **PSE**: pyschological stress evaluator, developed by A. Bell,Verimetrics (U.S. Patent by Bell and others, 1976).
- **CVSA**: Computerized voice stress analyzer, National Inst. Truth Ver., C. Humble.
- **Lantern**: Diogenes Group
- **Truster**: Makh-Shevet, Isreal company.
- Several low cost voice stress analyzer kits

Although the details by which these methods operate are not clearly described in their literature, the claims of success are well documented *in the company literature.* Most, if not all, of these methods focus on some aspect of assessing the presence of microtremors which are expected to be present when a speaker is under neutral/calm speaking conditions. These microtremors are expected to be reduced when a speaker is under stress. The results from our study here cannot prove or disprove the commercial claims. However, our evaluations using various linear and nonlinear based excitation features suggest that various types of emotion/stress can be detected in some individuals. The reliability will depend on the available training data for the classifier, and we expect that stress classification performance should be more successful if there is a means of "training" the system for a given speaker in similar conditions. Some of the claims made by these manufacturers have no basis, or are so extreme that they go against basic speech science. The Truster web-page states that their system will be able to determine deception even if the speaker is under different levels/types of emotion. Such a claim has no scientific merit, since it is not possible to cleanly separate the excitation signal into component dues to emotion and those due to deception.

More recent algorithms for voice stress analysis have been proposed using digital speech processing techniques, some of which suggest alternative excitation methods which offer the promise of better system integration within speech/speaker recognition or voice equipment for communications scenarios.

While research and progress have been made in the areas of stress classification and assessment, a number of important research areas require further investigation. Here, we briefly consider four points. First, in order to perform stress classification or assessment, two anchor models are needed (one for neutral and one for stress). These models should be trained using speech obtained from the actual stressful environments in which we wish to assess operators (i.e., aircraft pilot recordings if pilots are to be assessed; subject interviews in law enforcement). The type of stress which is displayed in one setting (aircraft cockpit), may not reflect the same conditions experienced in another (law enforcement questioning session). Second, further research is needed to assess the consistency of stress assessment/classification for a given speaker and for unseen speakers (i.e., explore the impact of using other training data to assess new speakers). Commercial systems assume that the same feature will be effected by all speakers. There needs to be a way of determining if a stress classification algorithm/system would prove to be useful, or if the speaker is not a viable candidate for assessment. Third, there is clearly a range of emotions and psychological factors which all contribute to speaker 'stress.' In emergency scenarios a pilot may experience a combination of fear, anxiety, fatigue, etc. at the same time. A suspect under questioning would also display natural stress even if he were not guilty. The ability to classify/assess this mixture of speaker traits is important in determining the stress state of the speaker. Finally, there exists an unknown relationship between how computer based speech systems are able to classify stress and how humans perform stress classification. This operation is well documented in the field of speech quality assessment, where there exists scientifically recognized subjective tests, which are used to determine a degree of correlation with numerical objective measures. It would make sense to explore the field to determine if standardized tests exist or could be modified to subjectively determine stress state and level in speakers, and then apply either commercial systems or research based stress classification algorithms to determine their 'correlation' to correct stress detection. This issue is important in the collection of future databases so that better stress anchor models can be used with emerging speech technology. From the research conducted here, it is suggested that speakers often vary how they convey stress in their speech, and that several speech features may be needed to

capture the subtle differences in how speakers convey their stress state in different voice communications scenarios.

Software implementations of the features presented in this report will be supplied directly to the sponsor. This will include algorithms coded in Matlab and *C* of linear features, CVSA, and the TEO-CB-Auto-Env measure.

9 References

Speech Communication (**1996**) Special Issue on Speech under "Stress", **20**(1-2):3-173, Nov.

Arslan, L. (**1996**)."Foreign Accent Classification", Ph.D. Thesis, Robust Speech Processing Laboratory, Duke University, July.

Baber, C., Mellor, B., Graham, R., Noyes J.M., Tunley, C. (**1996**)."Workload and the Use of Automatic speech recognition: The Effects of Time and Resource Demands," *Speech Communication*, **20**(1--2) 37-54.

Bachrach, A.J.\ (**1979**)."Speech and its Potential for Stress Monitoring: Monitoring Vital Signs in the Diver," Naval Medical Research Institute TECHNICAL REPORT, Aug., 78-93.

Barnwell, T.P. (**1971**)."An Algorithm for Segment Durations in a Reading Machine Context," Final Technical Report 479, Research Laboratory of Electronics, Mass. Inst. Of Tech., Cambridge, MA.

Bell, A.D., Ford, W.H., McQuiston, C.R., "Physiological Response Analysis Method and Apparatus," U.S. Patent No. 3,971,034, July 20, 1976.

Bond, Z.S. and Moore, T.J. (**1990**). "A note on Loud and Lombard Speech," *ICSLP-90*, Kobe, Japan, 969-972.

Cairns, D.A. and Hansen, J.H.L. (**1994a**). "Nonlinear Analysis and Detection of Speech Under Stressed Conditions," J. Acoust. Soc. Am. **96**(6) 3392-3400.

Cairns, D.A. and Hansen, J.H.L. (**1994b**). "Nonlinear Speech Analysis using the Teager Energy Operator with Application to Speech Classification under Stress," *ICSLP-94*, **II**(3):1035-1038, Yokohama, Japan, Sept.

Chen, Y. (**1987**)."Cepstral Domain Stress Compensation for Robust Speech Recognition," *IEEE ICASSP-87*, Dallas, TX, 717-720.

Cestaro, V. L., "A Comparison between the Decision Accuracy Rates Obtained Using the Polygraph Instrument and the Computer Voice Stress Analyzer (CVSA) in the Absence of Jeopardy," Tech Report, DoD Polygraph Inst., August 1995.

Creelman, C.D. (**1962**). "Human Discrimination of Auditory Duration," J. Acoust. Soc. Am. **34**(5) 582-593.

Darby, J.K. (**1981**).*Speech Evaluation in Psychiatry*, Grune & Stratton, New York, New York.

Davis, S., and Mermelstein, P. (**1980**).Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, and Signal Process.*, **ASSP-28**(4) Aug., 357--366.

Deller, J.R., Hansen, J.H.L., and Proakis, J.G. (**1999**)."Discrete-Time Processing of Speech Signals," 2nd Ed., IEEE Press, New York, NY.

Flack, M. (**1918**)."Flying Stress," London: Medical Research Committee.

Folds, D.J., Gerth, J.M., Engelman, W.R. (**1986**)."Enhancement of Human Performance in Manual Target Acquisition and Tracking," Final Tech. Report USAFASM-TR-86-18, USAF School of Aerospace Med., Brooks AFB, TX.

Folds, D.J. (**1987**)."Response Organization and Time-Sharing in Dual-Task Performance," Ph.D. Thesis, School of Psychology, Georgia Inst. of Tech., Atlanta, GA.

Fry, D.B. (**1955**)."Duration and Intensity as Physical Correlates of Linguistic Stress," J. Acoust. Soc. Am. **27**(4) 765-768.

Gardner, M.B. (**1966**)."Effect of Noise System Gain, and Assigned Task on Talking Levels in Loudspeaker Communication," J. Acoust. Soc. Am. **40**(5) 955-965.

Goldberger, L., Breznitz, S. (**1982**).*Handbook of Stress: Theoretical & Clinical Aspects*, Free Press, Macmillan Pub., New York, New York.

Gong, Y. (**1995**)."Speech recognition in noisy environments: A survey," m *Speech Communication*, **16**:261--291.

Hanley, C.N., Harvey, D.G. (**1965**)."Quantifying the Lombard Effect," J. of Hearing & Speech Disorders, **30**:274-277.

Hansen, J.H.L. (**1988**). "Analysis and Compensation of Stressed and Noisy Speech with Application to Robust Automatic Recognition," Ph.D. Thesis, School of Electrical Engineering, Georgia Inst. of Tech., Atlanta, GA.

Hansen, J.H.L. (**1989**). "Evaluation of Acoustic Correlates of Speech Under Stress for Robust Speech Recognition," IEEE Proc. 15th Northeast Bioengineering Conf., Boston, Mass., 31-32.

Hansen, J.H.L. (**1993**)."Adaptive Source Generator Compensation and Enhancement for Speech Recognition in Noisy Stressful Environments," ICASSP-93, Minn., MN, 95-98.

Hansen, J.H.L. (**1994**). "Morphological Constrained Enhancement with Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect," IEEE Trans. on Speech & Audio Proc **2**(4):598-614.

Hansen, J.H.L. (**1996**). "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition, *Speech Communications: Special Issue on Speech Under Stress*, **20**(1-2):151--173.

Hansen, J.H.L. (**1998a**)."Analysis of Acoustic Correlates of Speech Under Stress.Part I: Fundamental Frequency, Duration, and Intensity Effects," submitted Oct. 20, 1998 to *J. Acoust.* Soc. *Am.*

Hansen, J.H.L. (**1998b**)."Analysis of Acoustic Correlates of Speech Under Stress.Part II: Glottal Source and Vocal Tract Spectral Effects," submitted Oct. 20, 1998 to *J. Acoust. Soc. Am.*

Hansen, J.H.L., Bria, O.N. (**1990**)."Lombard Effect Compensation for Robust Automatic Speech Recognition in Noise," *ICSLP-90*, Kobe, Japan, 1125-1128.

Hansen, J.H.L., Bou-Ghazale, S.E. (**1995**). "Duration and Spectral Based Stress Token Generation for Keyword Recognition Using Hidden Markov Models," *IEEE Trans. on Speech & Audio Proc.*, **3**(5), 415-421.

Hansen, J.H.L., Bou-Ghazale, S.E. (**1997**)."Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database," *EUROSPEECH-97*, **4**:1743-1746, Rhodes, Greece.

Hansen, J.H.L., Cairns, D.A. (**1995**)."ICARUS: Source generator based real-time recognition of speech in noisy stressful and Lombard effect environments," *Speech Communications*, **16**:391-422.

Hansen, J.H.L., Clements, M.A. (**1987**)."Evaluation of Speech under Stress and Emotional Conditions," Proc. J. Acoust. Soc. Am., **82**(Fall Sup.):S17, Nov.

Hansen, J.H.L. and Clements, M.A. (**1989**)." Stress and Noise Compensation Algorithms for Robust Automatic Speech Recognition," *IEEE ICASSP-89,* Glasgow, Scotland, U.K., 266-269.

Hansen, J.H.L. and Clements, M.A. (**1995**)."Stress and Noise Compensation Algorithms for Robust Automatic Speech Recognition," *IEEE Trans. on Speech & Audio Proc.*,} **3**(5):407-415.

Hansen, J.H.L., Gavidia-Ceballos, L., and Kaiser, J.F., (**1998**)."A nonlinear based speech feature analysis method with application to vocal fold pathology assessment," *IEEE Trans. On Biomedical Engineering*, **45**(3):300-313, March 1998.

Hansen, J.H.L., Mammone, R., Young, S. (**1994**)."Editorial for the SPECIAL ISSUE: Robust Speech Recognition," *IEEE Trans. on Speech & Audio Proc.*, **2**(4):549-550.

Hansen, J.H.L., South, A.J., Swail, C., Moore, R.K., Steeneken, H.J.M., Cupples, E.J., Anderson, T., Vloeberghs, C., Trancoso, I., Verlinde, P. (**1999**). *The Impact of Speech Under "Stress" on Military Speech Technology*, Final Technical Report, NATO AC/232/IST/TG-01, March.

Hansen, J.H.L., Womack, B. D. (**1996**). "Feature Analysis and Neural Network Based Classification of Speech Under Stress," *IEEE Trans. Speech Audio Proc.*, **4**(4):307-313.

Hanson, B.A., Applebaum, T.H. (**1990**)."Robust speaker-independent word recognition using static, dynamic and acceleration features: Experiments with Lombard and noisy speech," *IEEE ICASSP-90*, pp. 857- 860.

Haward, L. (**1975**) "Emotional Stress and Flying Efficiency," *AGARD Conf. Proceedings No.181*, North Atlantic Treaty Organization, C8-1, Oct.

Hecker, M.H.L., Stevens, K.N., von Bismarck, G., Williams, C.E. (**1968**) "Manifestations of Task-Induced Stress in the Acoustic Speech Signal," J. Acoust. Soc. Am **44**(4), 993-1001.

Hecker, M.H.L. (**1974**) "A Study of the Relationships Between Consonant-Vowel Ratios and Speaker Intelligibility," Ph.D Thesis, Stanford University, Palo Alto, CA.

Hess, W. (**1983**) *Pitch Determination of Speech Signals*, Springer Verlag, New York, NY.

Hicks, J.W., Hollien, H., (**1981a**)."The Reflection of Stress in Voice-1: Understanding the Basic Correlates," The 1981 Carnahan Conf. on Crime Countermeasures, 189-195.

Hollien, H., Hicks, J.W. (**1981b**)."The Reflection of Stress in Voice-2: The Special Case of Psychological Stress Evaluators," The 1981 Carnahan Conf. on Crime Countermeasures, May, 196-197.

Hollien, H., Majewski, W. (**1977**)."Speaker Identification by Long-Term Spectra Under Normal and Distorted Speech Conditions," *J. Acoust. Soc. Am.*, **62**(4):975-980.

Hollien, H., Majewski, W., Doherty, E.T. (**1982**)."Perceptual identification of voices under normal, stress and disguise speaking conditions," *J. Phonetics*, **10**:139-148.

House, A.S. (**1962**)."On Vowel Duration in English," J. Acoust. Soc. Am. **33**(9) 1174-1178.

Jelinek, J. (**1997**). *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, MA.

Jex, H.R. (**1979**)."A Proposed Set of Standardized Sub-Critical Tasks For Tracking Workload Calibration," in N. Moray, *Mental Workload: Its Theory and Measurement*, New York: Plenum Press, 179-188.

Junqua, J.C. (**1993**)."The Lombard reflex and its role on human listeners and automatic speech recognizers, *J. Acoust. Soc. Am.*, **93**(1):510-524.

Junqua, J.C. (**1996**). "The Influence of Acoustics on Speech Production: A Noise-induced Stress Phenomenon Known as Lombard Reflex," *Speech Comm.*, **20**(1-2):13-22.

Junqua, J.C., Fincke, S., Field, K. (**1999**)."The Lombard Effect: A reflex to better communicate with others in noise," *IEEE ICASSP-99*, **4**:2083-2086.

Kaiser, J.F. (**1990**)."On a Simple Algorithm to Calculate the 'Energy' of a Signal," *IEEE ICASSP-90*, pp. 381-384.

Kaiser, J.F. (**1993**)."Some Useful Properties of Teager's Energy Operator," *IEEE ICASSP-93*, **3**:149-152.

Klatt, D. (**1973**)."Interaction between two factors that influence vowel duration," J. Acoust. Soc. Am. **54**(4) 1102-4.

Klatt, D. (**1976**)."Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," J. Acoust. Soc. Am. **59**(5) 1208-21.

Kohler, K.J. (**1986**)."Invariance and Variability in Speech Timing: From Utterance to Segment in German," Chap. 13, *Invariance and Variability in Speech Processes*, ed.s J. Perkell & D. Klatt, Lawrence Erlbaum Ass., Hillsdale, NJ.

Kuroda, I., Fujiwara, O., Okamura, N., Utsuki, N. (**1976**)."Method for Determining Pilot Stress Through Analysis of Voice Communication," Aviation, Space, and Environmental Medicine, May, 528-533.

Levinson, S.E. (**1986**) "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, **1**:29-45.

Lieberman, P., Michaels, S. (**1962**). "Some Aspects of Fundamental Frequency and Envelope Amplitude as Related to the Emotional Content of Speech," J. Acoust. Soc. Am. **34**(7) 922-927.

Lippmann, R.P., Mack, M., Paul, D., (**1986**)."Multi-Style Training for Robust Speech Recognition Under Stress," Proc. J. Acoust. Soc. Am. 110th Meeting, QQ10.

Lippmann, R.P., Martin, E.A., Paul, D.B. (**1987**)."Multi-Style Training for Robust Isolated-Word Speech Recognition," *IEEE ICASSP-87*, Dallas, TX, 705-708.

Lively, S., Pisoni, D., van Summers, W., Bernacki, R. (**1993**)."Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences," J. Acoust. Soc. Am., **93**(5) 2962-2973.

Lombard, E. (**1911**)."Le Signe de l'Elevation de la Voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, **37**, 101-119.

Malkin, F.J., Christ, K.A. (**1985**)."Human Factors Engineering Assessment of Voice Technology for the Light Helicopter Family," U.S. Army Human Engineering Lab. Tech. Report, 1-20, June.

Maragos, P., Kaiser, J.F., and Quatieri, T.F. (**1993a**)."Amplitude and Frequency Demodulation Using Energy Operators," *IEEE Trans. on Signal Proc.*, **41**(4):1532-1550.

Maragos, P., Kaiser, J.F., and Quatieri, T.F. (**1993b**)."Energy Separation in Signal Modulations with Application to Speech Analysis," *IEEE Trans. on Signal Proc.*, **41**(10):3025-3051, Oct.

Martin, E.A., Lippmann, R.P., Paul, D.B., (**1987**)."Two-Stage Discriminant Analysis for Improved Isolated-Word Recognition," ICASSP-87, Dallas, TX, April, 713-716.

Murray, I.R., Arnott, J.L., Rohwer, E.A. (**1996**)."Emotional stress in synthetic speech: Progress and future directions," *Speech Communication*, **20**:85-92.

Murray, I.R., Baber C., South, A. (**1996**)."Towards a Definition and Working Model of Stress and Its Effects on Speech," *Speech Comm.*, **20** (1-2):3-12.

Nicholson, A.N., Hill, L.E., Borland, R.G., Krzanowski, W.J. (**1973**)."Influence of Workload on the Neurological State of the Pilot During the Approach and Landing," *Aerospace Medicine*, **44**(2) 146-152.

Ofuka, E., Valbret, H., Waterman, M., Campbell, N., Roach, P. (1994) "The role of F0 and duration in signaling affect in Japanese: anger, kindness, and politeness," *ICSLP-94*, **3**:1447-1450.

Paul, D.B. (**1987**)."A Speaker-Stress Resistant HMM Isolated Word Recognizer," *IEEE ICASSP-87*, pp. 713-716.

Pearsons, K.S., Bennett, R.L., Fidell, S. (**1977**)."Speech Levels in Various Noise Environments," Office of Health and Ecological Effects, Report No. EPA-600/1-77-025.

Peckham, J.B. (**1979**)."A Device For Tracking The Fundamental Frequency of Speech and its Application in the Assessment of 'Strain' in Pilots and Air Traffic Controllers," Tech. Report 79056, Royal Aircraft Est., May, 1-55, 1979.

Perkell, J.S., Klatt, D.H. (ed.'s) (**1986**).*Invariance and Variability in Speech Processes*, Lawerance Erlbaum Ass., Hillsdale, N.J.

Pickett, J.M. (**1980**).*The Sound of Speech Communication*, University Park Press, Baltimore, Maryland.

Pisoni, D.B., Bernacki, R.H., Nusbaum, H.C., Yuchtman, M. (**1985**)."Some Acoustic-Phonetic Correlates of Speech Produced in Noise," ICASSP-85, Tampa, Fl, 41.10.1-4.

Poock, G.K., Armstrong, J.W. (**1981**)."Effect of Operator Mental Loading on Voice Recognition Systems Performance," Naval Postgraduate School Tech. Report, Aug.

Poock, G.K., Armstrong, J.W. (**1981**)."Effect of Task Duration on Voice Recognition System Performance," Naval Postgraduate School Tech. Report, Sept.

Rabiner, L., Juang, B.-H. (**1993**).*Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

Ramig, L. (**1992**)."The role of phonation in speech intelligibility: A review and preliminary data from patients with Parkingson's disease," in Kent, R.D. *Intelligibility in Speech Disorders*, John Benjamins Pub. Co., Philadephia, PA, 1992.

Rajasekaran, P.K., Doddington, G.R., Picone, J.W. (**1986**)."Recognition of Speech Under Stress and In Noise," *IEEE ICASSP-86,*} Tokyo, Japan, 733-736.

Reed, L. (**1985**)."Military Applications of Voice Technology," *Speech Technology*, Feb., 42-50.

Rostolland, D. (**1982**)."Acoustic Features of Shouted Voice Part I," *Acustica*, **50** 118-125.

Rostolland, D. (**1982**)."Phonetic Structure of Shouted Voice Part II," *Acustica*, **51** 80-89.

Ruiz, R., Absil, E., Harmegnies, B., Legros, C., Poch, D. (**1996**)."Time and spectrum-related variabilities in stressed speech under laboratory and real conditions, *Speech Communication*, **20**:111-130.

Shipp, T., Izdebski, K., "Current Evidence for the Existance of Laryngeal Macro and Microtremor," Journal of Forensic Sciences, **26**(3):501-505, July 1981.

Simonov, P.V., Frolov, M.V. (**1977**)."Analysis of the Human Voice as a Method of Controlling Emotional State: Achievements and Goals," Aviation, Space, & Environmental Sciences, Jan., 23-25.

Simpson, C.A. (**1985**). "Speech Variability Effects on Recognition Accuracy Associated With Concurrent Task Performance by Pilots," Psycho-Linguistic Research Associates, Technical Report, April.

Stanton, B.J., Jamieson, L.H. and Allen, G.D. (**1988**)."Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions," *IEEE ICASSP-88: Inter. Conf. on Acoust., Speech, Sig. Proc.*, pp.331-334.

Stanton, B.J., Jamieson, L.H. and Allen, G.D. (**1989**)."Robust recognition of loud and lombard speech in the fighter cockpit environment," *IEEE ICASSP-89*, pp. 675-678.

Steeneken, H.J.M., Hansen, J.H.L. (**1999**)."Speech under Stress Conditions: Overview of the Effect of Speech Production and on System Performance," *IEEE ICASSP-99*, **4**:2079-2082.

Summers, W.V., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I., Stokes, M.A. (**1988**)."Effects of Noise on speech production: Acoustic and Phonetic Analysis," J. Acoust. So. Am. **84**(3):917-928.

Sweet, W. (**1995**). "The Glass Cockpit," *IEEE Spectrum*, September 1995, pp. 30-38.

Streeter, L.A., Macdonald, N.H., Apple, W., Krauss, R.M., Galotti, K.M. (**1983**)."Acoustic and Perceptual Indicators of Emotional Stress," J. Acoust. So. Am. **73**(4):1354-1360.

Takizawa, Y., Hamada, M. (**1990**). "Lombard speech recognition by formant-frequency-shifter LPC cepstrum," *ICSLP-90: Inter. Conf. on Spoken Lang. Proc.*, pp. 293-296.

Teager, H. (**1980**)."Some Observations on Oral Air Flow During Phonation", *IEEE Trans. Acoust., Speech, Signal Proc.*, **28**(5):599-601, Oct.

Teager, H., Teager, S. (**1983**)."A Phenomenological Model for Vowel Production in the Vocal Tract," in *Speech Science: Recent Advances*, edited by R.G. Daniloff (College-Hill, San Diego), pp.73-109.

Thomson, D.L., Chengalvarayan, R., "Use of Periodicity and Jitter as Speech Recognition Features" *Proc. IEEE ICASSP'98*, Seattle, Washington, May 1998.

Umeda, N. (**1975**)."Vowel duration in American English," J. Acoust. Soc. Am. **58**(2):434-445.

Umeda, N. (**1977**)."Consonant duration in American English," J. Acoust. Soc. Am. **61**(3) 846-858.

van Santen, J.P. (**1995**)."Computation of timing in text-to-speech synthesis," in *Speech Coding and Synthesis*, Kleijn and Paliwal, eds., Elsevier N.Y.

van Santen, J.P., Hirschberg, J. (**1994**)."Segmental effects on timing and height of pitch contours," *ICSLP-94*, **2**:719-722.

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, **25**(1):174-188, Jan. 1980.

Wang, X., Pols, L.C., ten Bosch, L.F. (**1996**)."Analysis of context-dependent segmental duration for automatic speech recognition," *ICSLP-96*, **2**:1181-84.

Whitmore, J., Fisher, S. (**1996**)."Speech during sustained operations," *Speech Communication*, **20**:55-70.

Williams, C. E., and Stevens, K. N. (**1969**). "On Determining the Emotional State of Pilots During Flight: An Exploratory Study," Aerospace Medicine, **40** 1369-1372.

Williams, C.E., and Stevens, K.N. (**1972**). "Emotions and Speech: Some Acoustic Correlates," J. Acoust. Soc. Am., **52**(4) 1238-1250.

Womack B.D., Hansen, J.H.L. (**1996**)."Classification of Speech under Stress Using Target Driven Features," *Speech Comm.*, **20**(1-2):131-150, Nov.

Yost, W.A. (**1994**).*Fundamentals of Hearing*, 3rd Edition, Academic Press, San Diego, CA., pp. 153-167.

Zhou, G., Hansen, J.H.L., Kaiser, J.F. (**1998a**)."Classification of Speech under Stress Based on Features from the Nonlinear Teager Energy Operator," *IEEE ICASSP-98*, **1**:549-552, Seattle, WA.

Zhou, G., Hansen, J.H.L., Kaiser, J.F. (**1998b**)."Linear and Nonlinear Speech Feature Analysis for Stress Classification," *ICSLP-98*, **3**:883-886, Sydney, Australia.

MISSION
OF
AFRL/INFORMATION DIRECTORATE (IF)

The advancement and application of Information Systems Science and Technology to meet Air Force unique requirements for Information Dominance and its transition to aerospace systems to meet Air Force needs.