NASA Contractor Report 4249

Voice Measures of Workload in the Advanced Flight Deck

Sid J. Schneider, Murray Alpert,
and Richard O'Donnell
Behavioral Health Systems, Inc.
Ossining, New York

Prepared for
Langley Research Center
under Contract NAS1-18278

National Aeronautics and
Space Administration
Office of Management
Scientific and Technical
Information Division

1989

TABLE OF CONTENTS

ABSTRACT

Voice samples were obtained from 14 male subjects under high and low workload conditions. The subjects said, "Triangle please stop turning now" when one of two triangles on a computer monitor began to rotate, at approximately 25 sec intervals. The subjects simultaneously performed a secondary task, the Continuous Performance Test (CPT), which was presented at the center of the computer monitor. The presentation rate of the CPT was continually adjusted so that the error rate remained at .6 in the high workload condition and .2 in the low workload condition. Two trials in each workload condition were run.

Each utterance, "Triangle please stop turning now," was divided into peaks, or points of high amplitude relative to surrounding points. The frequency, amplitude and duration of each peak was determined. Then, the means and variances of these measures were calculated for each utterance. The grand means of the means and variances of the utterances were calculated for each trial.

Analyses of the grand means revealed a trend for the high workload condition to bring about higher amplitudes and frequencies. However, the trend did not reach statistical significance in the present data, or in data from other laboratories. Analyses of the present data revealed that there was much individual difference in how workload affected the grand means of the acoustical measures.

Further analyses revealed statistically significant temporal patterns in the data. In the high workload condition, frequency and amplitude fell over the course of both runs. In the second run, these measures never recovered to their earlier, higher levels. They continued to fall from relatively low levels in the second run, though frequency later increased somewhat.

By contrast, amplitude did recover to its earlier, higher level at the start of the second run of the low workload condition. Also, amplitude and frequency did not display as sharp a dropoff over time as they did in the high workload condition.

The results suggest that the grand means of acoustical measures may not be reliable indicators of high workload situations on the flight deck. High workload conditions are better revealed by their effects on the voice over time. Aircrews in the advanced flight deck will be voicing short, imperative sentences repeatedly. A drop in the energy of the voice, as reflected in amplitude and frequency, followed by the failure to achieve old energy levels after rest periods, can signal that the demands of the situation are taking a toll on the speaker. This kind of measurement would be relatively unaffected by individual differences in acoustical measures.

INTRODUCTION

This purpose of this study was to determine the relationship between the acoustical properties of the voice and the workload demands placed upon the speaker. The study was designed to be applicable to the advanced flight deck, in which the flight crew executes many tasks by speaking commands. Voice recognition devices on the advanced flight deck determine which command has been issued, and take the appropriate actions. Voice recognition devices allow the flight crew to perform many tasks by speaking, rather than by manually adjusting instruments. In this way, fewer tasks involve the flight crew's hands and eyes, and more involve their ears and voice. Since the hands and eyes are heavily used in a flight situation, this redistribution of tasks to the ears and voice may reduce the competition among tasks for identical channels. By keeping the manual and visual channels from becoming overloaded, the advanced flight deck may prevent decrements in pilot performance.

The development of the advanced flight deck may also provide a new direction to research on workload. At present, workload is assessed with many measures, including subjective reports, and electrocardiographic, electroencephalographic, and electrodermal records. The advanced flight deck allows researchers to obtain many voice samples from pilots over the course of a flight. Obtaining voice samples is unobtrusive and well tolerated; it does not entail placing electrodes on the body, or filling out questionnaires. One problem with using the voice to measure workload is simply the lack of knowledge concerning how best to do it. It must be determined which computer techniques for analyzing the voice best bring out those acoustic parameters that are affected by workload. It must also be determined specifically how increases in workload affect those parameters.

In the present study, subjects were asked to perform two simultaneous tasks. One task elicited speech samples. The subjects spoke short, imperative statements, as pilots in the advanced flight deck would do. The second task was a "loading" task, designed to change the workload demand placed upon the subject. The time pressure of the loading task was great in a high workload condition, smaller in a low workload condition. The voice samples were subjected to computer analysis, so that the frequency, volume, tempo, and monotony of the voice could be quantified.

The computer apparatus used in this analysis has been employed in previous studies of the mental state of psychiatric patients. That research has uncovered several acoustical parameters of the voice that are sensitive to changes in depression and other psychopathological states. The present research was to uncover which acoustical parameters were sensitive to workload.

VOICE INTERACTION TECHNOLOGY IN THE ADVANCED FLIGHT DECK

The technology for voice recognition and synthesis will soon be accurate and economical enough to find wide applications. Already, several inexpensive systems that can recognize words in natural human speech are commercially available. Voice recognition and synthesis devices will certainly play increasing roles in piloting activities in many kinds of aircraft. These devices may reduce workload and save time in the cockpit, with only minimal disruption of communication and other voice activities of the pilot.

Present flight deck environments place demands on many of the pilot's senses. Visual inputs include the instruments, charts, maps, the aircraft, and the surroundings. Auditory inputs include radios, engine noise, and warnings. Vestibular input helps to monitor aircraft orientation. While these inputs are monitored, manual responses to control the aircraft must be performed. Competition between inputs in the same modality can result in a greater reduction in accuracy than would competition between inputs in different modalities. Voice recognition technology can reduce competition for manual and visual processing channels (Porubcansky, 1985). For example, it is often necessary for pilots simultaneously to watch aircraft instruments and to watch events outside the aircraft. If the task of watching the instruments could be converted to the vocal and auditory modalities through voice recognition and synthesis devices, the competing tasks would not both be visual tasks. Workload would diminish, and the pilot would be less taxed.

Lea (1983) found that voice recognition devices allowed users to enter data more rapidly than they could manage manually. Moreover, users who never used a voice system before required relatively little time to learn to enter data at a fast rate using a voice system; more time was needed to train users of a manual system. Therefore, even without considering the competition among tasks on a flight deck, voice recognition systems may be less demanding upon their users than are manual systems. North and Bergeron (1984) argued that the reduction in pilot workload might prevent poor or late recognition of hazardous situations, flight path deviation, inaccurate adjustments of equipment, or incorrect procedures.

In sum, voice recognition and synthesis devices could redistribute piloting tasks so that the hands and eyes would not have to participate in as many of them. The pilot would gather data and perform piloting tasks by speaking commands and listening for responses and acknowledgments.

North and Bergeron (1984) surveyed pilots and engineers about which specific flight deck activities performed by the pilots of several kinds of instrument flight rules (IFR) aircraft could most benefit from voice recognition and synthesis technology. They found that the activities that could benefit

most from voice recognition technology, listed from the greatest benefit to the least benefit, were:

1. Interfacing with uplinked data (i.e., requesting airport and weather data, runway conditions and closures, and air traffic controller messages like altitude assignments).
2. Reprogramming waypoint.
3. Retrieving performance charts.
4. Retrieving approach charts.
5. Retrieving emergency checklists.
6. Setting up a course intercept.
7. Retrieving routine checklists.
8. Interacting with electronic map.
9. Requesting present location.
10. Tuning and resetting very high frequency omnidirectional range/distance measuring equipment (VOR/DME) facility.

Activities that would benefit to a lesser extent from voice recognition technology, according to the pilots and engineers, included: tuning the communications radios, setting the pressure altimeter, setting and changing the autopilot heading, altitude and air speed, and setting the heading bug, directional gyro, transponder code, and rate of descent and climb.

North and Bergeron (1984) also found many potential applications for voice synthesis technology, such as automatic reports of direction to the airport facility, and announcements of fuel level, altitudes, airspeed, and system warnings.

In a different review performed for NASA by the Boeing Commercial Airplane Company, White and Parks (1985) similarly concluded that voice recognition technology had numerous applications in the cockpit. For communications, the technology could be used for selecting communications modes, volume control, entering frequencies, and tuning the radio. In navigation, the applications include entering navigational radio frequencies. Voice recognition could be used to select the position for flaps, speed brake and trims, selecting autopilot and fuel management system modes, and entering data to the autopilot and throttle. Aircraft subsystems like hydraulics, air conditioning, anti-ice, rain and fire protection, and landing gear could be controlled by voice commands. For the flight instruments, voice recognition technology could be applied to select speed and height bugs, and to enter the barometric pressure.

The vocabulary and syntax used in voice commands on the advanced flight deck need to be highly constrained in order to minimize recognition errors. White and Parks (1985) estimated that a vocabulary of about 50 words could handle all commands that the devices could receive. Similar sounding words would be eliminated from the vocabulary.

North and Bergeron (1984) wrote that the greatest human

factors problem for cockpit voice recognition devices is keeping the pilots from forgetting what they must say for the system to function correctly. One technique to minimize forgetting has been to make the vocal commands analogous to the actions of turning dials and pressing keys that the commands replace. However, this technique led to lengthy, awkward commands.

A more desirable approach has been to use a "task oriented grammar" with its own syntax structure designed specifically for voice recognition devices. All commands in task oriented grammar are short, and take a specific form: verb (e.g., "set"), followed by noun (e.g., "VOR"), followed by modifier (e.g., number 2), followed by data (e.g., "to 110.0"). This grammar always results in short commands (e.g., "Set VOR number 2 to 110.0"). Its one drawback is its difficulty handling slight deviations from its structure.

Examples of commands issued using "task oriented grammar" are:

Interacting with uplinked data: "Report winds aloft at JFK"
Requesting present location: "Report present location"
Requesting direction to facility: "Where is Des Moines?"
Tuning communications radios: "Set COM 1 to Minneapolis"
Setting autopilot: "Set autopilot heading to three-three-five"
Engaging and releasing autopilot: "Engage autopilot"
Setting transponder code: "Set transponder to four-four-five-two"
Switching fuel tanks: "Switch tanks"
Controlling internal lighting: "Set overhead to medium"

The short, imperative commands of task oriented grammar require the pilot to remember only a small vocabulary and a simple syntax structure. Test versions of the advanced flight deck have employed forms of task oriented grammar.

North and Bergeron (1984) proposed a test system for laboratory simulations of a flight deck that used voice synthesis and recognition. Two microcomputers would be used. The first would simply run one of the flight simulation programs, such as the simulation of the Cessna 182 distributed by Sublogic and Microsoft, or the DC-10 simulator by Michtron. The second microcomputer would communicate with the first, and use the speech recognition and synthesis hardware. When the operator stated, for example, "Tune VOR to Minneapolis," the voice recognition microcomputer would appropriately set the VOR that is simulated in the program running on the first computer. North and Bergeron (1984) wrote that voice recognition tasks that could easily be simulated in this test system include changing VOR and communication radio frequencies, setting of direction gyro (DG), pressure altimeter, and transponder, and setting a radial into the navigation system. Voice synthesis could be used in this test system for altitude and fuel level warnings, and a read-through

of checklists.

The flight deck simulation that we used in this study resembles North's proposed system in that the voice of the subjects controlled actions that occurred on a computer screen. The subjects had to say a short, imperative sentence, similar in structure to those used on an advanced flight deck. The subjects had to simultaneously perform a task which was designed to vary the workload experienced by the subject. The speech samples that were collected during the tasks were computer analyzed, in order to quantify the acoustical properties of the speech, and determine which acoustical measures reflected the workload imposed upon the subjects.

VOICE ACOUSTICAL ANALYSIS SYSTEM

The apparatus to be described was originally designed to provide objective and quantitative measures of psychiatrically relevant variations in feeling states such as depression, mania, or the flat affect of schizophrenia (Alpert, Homel, Merewether, Martz, & Lomask, 1986). Further work with the apparatus revealed that the irritable, easily aroused speech pattern of the "Type A" individual who is at risk of developing coronary disease was also reflected in the voice measures (Alpert, 1982). Other research explored the effects of psychotropic medications on psychiatric states. Empirical studies have targeted the most discriminating acoustic parameters for each of these variables.

Signal analysis is done by a hybrid analogue/digital device which provides information about such prosodic voice variables as fundamental frequency (pitch), amplitude (loudness for each syllable) and the duration of utterances and latencies, and the variances of these measures.

Real time analysis provides on line information about the speaker's state. Analogue signal processing articulates the prosodic features of interest, eliminating most of the complexity of the speech waveform which is irrelevant to the task of measuring workload. Thus, while extraction of phonetic information would require a bandwidth of up to about 3 or 4 KHz and a sampling rate of up to ten times the bandwidth, for the present application, since speech syllable production rate is under five per second, digital sampling of 100 per second is more than adequate. Thus, the analogue conditioning greatly reduces the load on digital processing while enhancing the ability to identify the prosodic features of interest by eliminating irrelevant noise. Diagram 1 shows the general physical layout of the system. It consists of three main components: 1) the signal capture equipment consisting of head worn microphones and a good quality stereo cassette tape deck, and a calibrator; 2) an analogue processing stage and a microcomputer equipped with an analog to digital conversion system (IBM PC/AT computer with a Tecmar TM-AD212 analogue to digital [A/D] converter); and 3) a multifunction analogue signal processing unit. The analogue

computer unit provides the circuitry for filtering and transforming the speech signals prior to digital analysis. The raw AC signal that comes from the tapedeck is first passed through a bandpass filter (6 dB/octave roll off) in order to restrict the signal to a range around the speaker's fundamental frequency, thus eliminating harmonic overtones. The range between the filters is adjusted for the particular voice; for example, it is usually set between 80 to 140 hertz for male voices and between 120 and 300 hertz for female voices.

Once filtered, the speaker's signal is then split into two parallel lines which are analyzed separately, one channel for frequency information, and one for amplitude information. The frequency signal goes through a frequency to voltage converter which outputs 1 volt for each 50 hertz of signal; this signal then goes to one of the channels on the A/D converter board with a resolution of 200 counts per volt. The resulting resolution is 4 counts per hertz. The signal on the amplitude line is first passed through an attenuator, to match the dynamic range of the rectifier, then full wave rectified, and finally demodulated to produce a voltage analogue that approximates speech syllables that then goes to another channel on the A/D converter board.

Diagram 2 gives some idea of the way the amplitude signal would look, if one were to do an oscillographic tracing of it, at the point of input in the A/D converter in the PC/AT. In the software there is a log lookup table so that the variation in voltage and frequency across time is made proportional to the logarithm of the amplitude and frequency of the voice.

An utterance is defined as an amplitude which is above some threshold of background noise for at least 100 msecs or more; a gap as an amplitude that goes below threshold for at least 200 msecs; and a peak as a point of maximum amplitude relative to the values of amplitude immediately preceding and following that point. A calibration reference signal of known amplitude and frequency is recorded on the subject's channel. Since the subject uses a head held microphone of known output, the use of a calibration signal permits a usable estimate of the absolute voice level despite adjustments and variations in recording and playback amplifiers.

The software was designed to measure the following prosodic features of speech:

1) Number of utterances, gaps, and peaks.

2) Mean and variance of the time durations of peaks (syllables), utterances, and pauses.

3) Mean and variance of: i) the natural logarithm of the amplitude of peaks (loudness) as well as, ii) the natural logarithm of the frequencies (pitch) corresponding to those peaks.

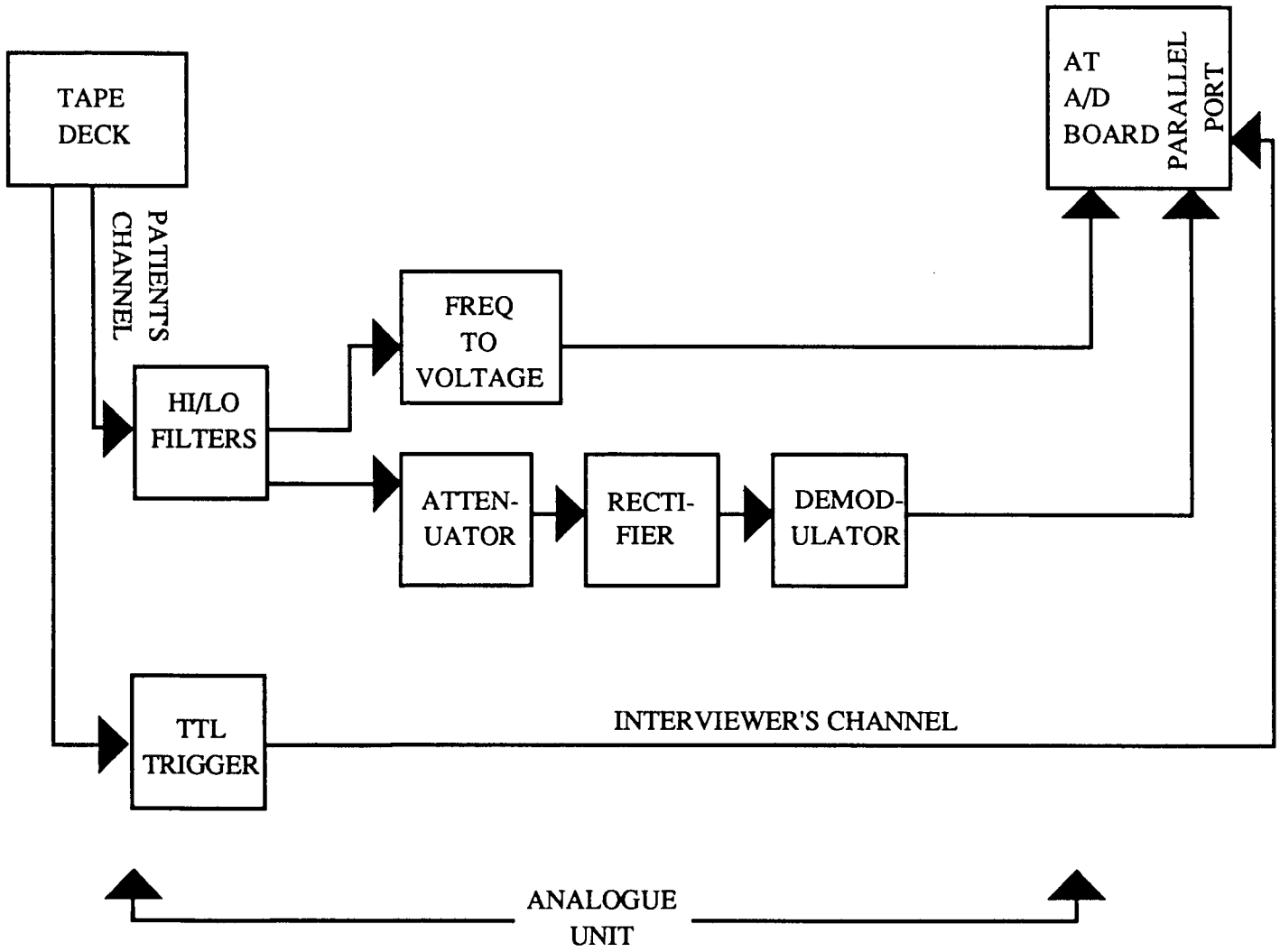
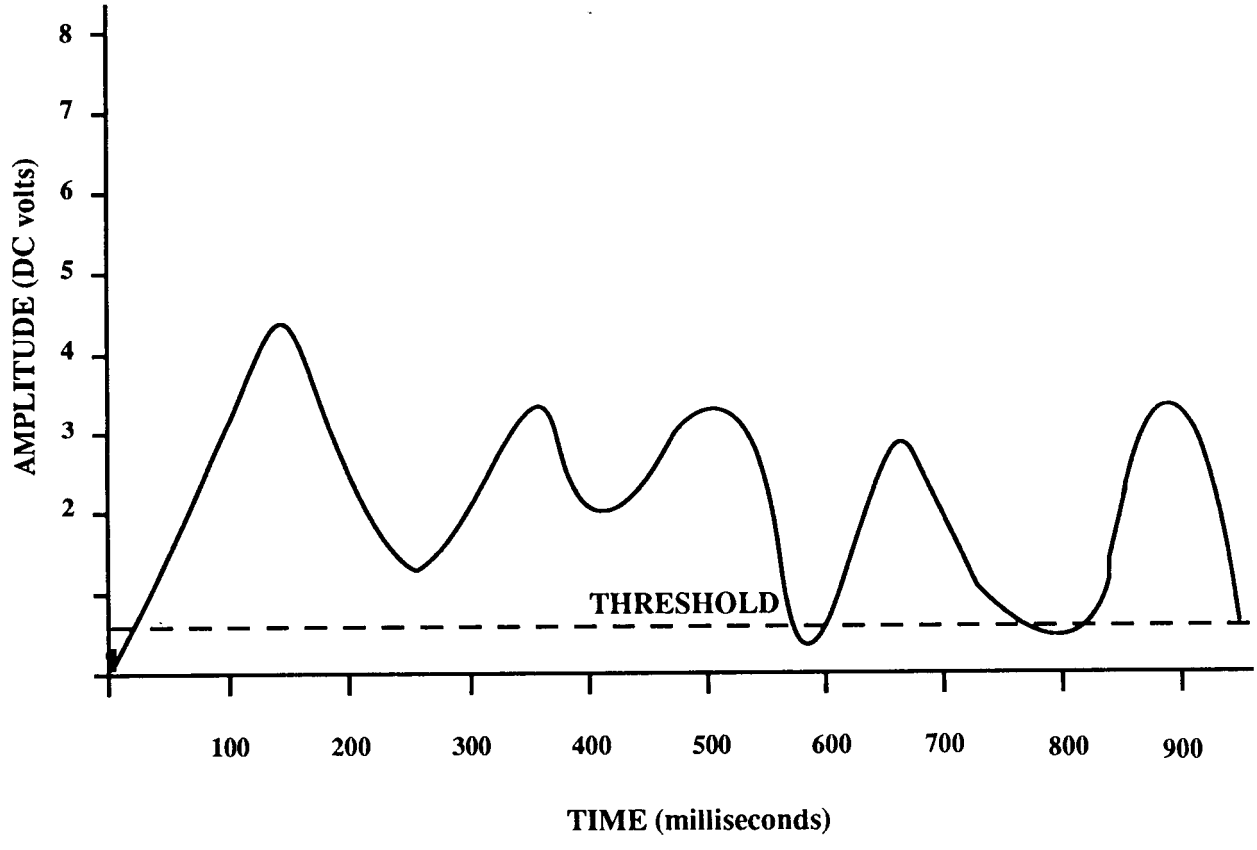

TAPE DECK
PATIENTS CHANNEL
HI/LO FILTERS
FREQ TO VOLTAGE
ATTEN-UATOR
RECTI-FIER
DEMOD-ULATOR
AT A/D BOARD
PARALLEL PORT
TTL TRIGGER
INTERVIEWER'S CHANNEL
ANALOGUE UNIT

DIAGRAM 1



AMPLITUDE SIGNAL AS A FUNCTION OF TIME

Diagram 2

It is possible to remove the effects of other speakers. Their speech, recorded on the second channel of the stereo deck, can be sent to a separate channel of the analogue computer, which detects the presence of a signal and sends a TTL signal, detected by the software, which suspends the analysis until the TTL signal is removed. In this way, the speech that is analyzed could be uncontaminated by other speakers and noises that may occur in an aircrew operational setting.

RESULTS OF PREVIOUS STUDIES

The apparatus for analysis of the human voice has been employed in a series of clinical studies at the Millhauser Laboratories for Research in Psychiatry and the Behavioral Sciences at New York University Medical Center. Several studies by Murray Alpert suggest that data from the apparatus are reproducible, highly precise, and useful in a clinical setting. For example, the apparatus provides an objective, reliable means of quantifying flat affect--the restricted emotions apparent in many schizophrenics--and distinguishing it from the clinically very similar presentation of patients with a retarded depression (Mayer, Alpert, Stestny, Perlick & Empfield, 1985). Flat affect is diagnostically important in schizophrenia. However, it is difficult to measure because other processes, such as psychomotor retardation, institutionalization, and drug side effects can mask it. Voice analysis provides a way to quantify flat affect in schizophrenics on the basis of diminished inflection (variation in frequency) and diminished dynamics (variation in volume). In depressives, the mood disturbances tends to be shown by long pauses and speech dynamics (Alpert, 1986). Measurement of flat affect using this apparatus compared favorably to clinical ratings made by highly skilled attending psychiatrists in evaluating and predicting patient behaviors (Alpert and Anderson, 1977).

Acoustic analysis had permitted the articulation of processes that are frequently confounded clinically and conceptually. Thus, it has become possible to distinguish affects from moods. Affects are encoded in voice emphasis and are measurable in the speaker's emphasis pattern and other acoustic parameters, such as average amplitude and duration of utterances and pauses. Affects reflect momentary feelings of which the speaker will not be aware. Moods, on the other hand, are encoded in temporal patterns of utterances and pauses and have much slower more enduring phases. Moods are associated with subjective feelings, like sadness and joy. They are revealed in the duration of pauses and utterances. It is important to distinguish both of these processes from emotions, like anger or fear, which are detectable in voice because they disrupt normal speech patterns. If a subject is emotionally aroused, the arousal affects physiological mechanisms important for speech. For example, changes in respiration will affect speech energetics, and changes in muscle tone will alter the overtone

structure and the speaker's voice quality (Alpert, 1982). These changes may be reflected in vocal alterations which last several seconds or minutes.

These insights into the separation of different feeling states grew out of studies with a variety of patient populations, treatment paradigms, and experimental procedures for producing emotional arousal in normals, such as having the subject lie or applying mildly aversive stimuli. It is noteworthy that these procedures can produce a vocal broadcasting of emotional arousal, in patients with depression or schizophrenia as well as in normal controls. The different feeling states appear to be controlled by different, complexly interacting brain mechanisms.

The apparatus has not, until now, been applied to the study of man-machine interactions. However, many of the psychological variables of interest in a clinical setting, like attention, arousal, and affect would also be of interest in human factors studies. The approach may well be appropriate to the study of multidimensional variables like workload.

WORKLOAD MEASUREMENT

The term "workload" refers to the cognitive demands placed upon an operator. Both the demands of the operator's tasks and the operator's ability to handle those demands figure into the workload level. Past writers have defined workload as "how busy is the operator," or "the aggregate of demands placed upon the operator," or "the descrepancy between system input and the system's capacity of handling that input." (Ogden, Levine, & Eisner, 1979). Workload is a multidimensional variable; it is a function of task demands, the operator's mental and physical capacity, work strategy, and skill level (Welford, 1978).

The workload experienced by flight crews has come under study because high workload levels may be associated with pilot error. There are diverse inputs impinging upon a pilot, who must perform three related tasks: monitor the inputs, interpret them, and intervene to change them when appropriate (Spettell & Liebert, 1986). Often, the conditions under which these operations are performed are stressful. Minimizing workload could mean optimizing pilot performance.

Measures of the workload imposed upon pilots must take into account the skill of the pilot, the number and the difficulty of the tasks before the pilot, and the psychological and physiological state of the pilot (Tole, Stephens, Vivaudou, Ephrath, & Young, 1983). Pilot workload has been modelled in the laboratory several ways. A computer program called POPCORN, for example, presents a set of simultaneous tasks to an operator, who is seated at a computer terminal. It is possible to manipulate workload in this program by changing the number and complexity of the tasks, as well as the penalties for errors and payoffs for correct responses (Hart, Battiste, & Lester, 1984). The program

is designed to reveal how manipulations of workload affect the operator's performance on the tasks.

Pilot workload has also been modelled in vehicle simulators. For example, Hicks and Wierwille (1979) manipulated workload by programming crosswinds in a vehicle simulator. Crosswinds near the front of the vehicle were said to create high workload; lower levels of workload were created by moving the crosswinds toward the rear. The authors found that an increase in workload affected several measures. First, the number of steering reversals increased. Steering reversals are corrections of oversteering and mark a lack of precision in steering control. As a result, the yaw and lateral deviations of the vehicle also increased.

In both POPCORN and vehicle simulators, the measure of workload is performance on the primary task. Investigators can observe the effects of workload on the performance scores in POPCORN, and on vehicle control in a simulator. However, there are other ways of assessing workload.

One straightforward way is to ask the operators to describe their level of workload. Several questionnaires, such as the "SWAT" or subjective workload assessment, have been designed to obtain operators' own opinions of their level of workload.

Many physiological measures have been employed to assess workload. Electroencephalographic, electrodermal, and electrocardiographic measures have been extensively studied (Rolfe & Lindsay, 1973; Wierwille, 1979).

The "secondary task" technique has also been extensively used to measure workload. This technique assesses workload by measuring how much spare mental capacity is available while the operator is busy with a primary task (Williges & Wierwille, 1979.) For example, if the primary task is operating a vehicle in a flight simulator, the secondary task might be a simultaneous digit reading task (Wierwille, Guttman, Hicks, & Muto, 1977). The subject's accuracy on the reading task would reveal how much capacity remains unused by the simulator task. A large amount of spare capacity would suggest that a low level of workload is imposed by the simulator task.

The "loading task" technique is similar to the secondary task technique. A loading task, however, is not used to measure workload; it is used to vary overall workload, in order to reveal the effects of workload upon performance on the primary task. In both the secondary and loading task techniques, there are two discrete and separate tasks, with a clear emphasis on one of the tasks. In the loading task technique, the difficulty of the secondary task is manipulated while performance on the primary task is observed (Ogden, Levine, & Eisner, 1979).

Loading tasks are useful for describing the relationship

between overall performance on the primary task and various levels of workload. Often, this relationship is complex, consisting of more than a linear decrement in performance as workload increases. Inverted-U or asymptotic relationships have been found in some situations (Ogden, Levine & Eisner, 1979).

The present study used the loading task paradigm to study the effects of workload upon acoustic measures of the voice. The primary task was an easy task which elicited speech from the subjects. At random intervals of about 25 seconds, the subjects had to say a short, imperative sentence in response to events on a computer screen. The simultaneous, loading task involved mental arithmetic. The time pressure of the loading task was varied in order to manipulate the demands placed on the subjects.

The loading task in this study was an "adaptive" task; that is, the time pressure was continually adjusted so that the subject's performance on the loading task remained stable. In the high workload condition, the error rate in the loading task was maintained at a high level; in the low workload condition, the error rate was maintained at a lower level. In this way, the voice samples elicited in the primary task could reveal the effect of increased workload on the acoustical properties of the voice.

VOICE MEASURES OF WORKLOAD

Everyday experience would suggest that the acoustical properties of the voice reflect the mental state of the speaker. The manner in which a person says, "Good morning," for instance, can lead others to conclude that the person is overworked, enthusiastic, tired, or worried. People constantly derive nonlinguistic information from the loudness, pitch, and tempo of other people's speech.

For some time, researchers have been seeking methods for analyzing the acoustical properties of the voice in order to monitor the mental state of the speaker. If such techniques were found to be reliable for a sizable number of individuals, they could be applied in aircraft environments, or other places where it is possible to collect voice samples.

The research on voice measures of mental state has been heterogeneous in many ways. Some researchers have focussed on the effects of mental state upon the fundamental frequency, while others have investigated various combinations of pitch, loudness, and tempo. Also, there have been varying techniques used to quantify these acoustic parameters. For example, the filtering and other treatment of the voice signal before analysis have varied among laboratories. The computer techniques used to derive the acoustical parameters have also varied. For example, some researchers use power spectrum analysis to derive the fundamental frequency, while others use zero-cross analysis.

Perhaps the two most important methodological discrepancies in the literature concern the nature of the speech that was analyzed, and the conceptualization of mental state. In some studies, natural, connected speech was analyzed; in others, speech samples were obtained by asking subjects to count; in others, the speech samples were very brief, disconnected utterances.

In much of the literature, researchers have written that they studied the effects of stress on the acoustic parameters of the voice. However, many of these researchers manipulated stress by changing the task demands placed on the subjects. The stess, it could be argued, was the result of an increase in workload. Other researchers studied the effects of stress by analyzing the voices of pilots during an impending potential catastrophe. That form of stress is probably attributable to fear, not workload.

Despite these methodological differences in the literature, some patterns in the findings are apparent.

One of the first studies to find a relationship between workload-induced stress and the voice was by Hecker, von Bismarck, and Williams (1967, 1968). Their subjects had to perform a series of tasks while uttering brief phrases. Some tasks were high stress tasks; they had to be performed under increased time pressure, which presumably increased the workload. The researchers, working before small computers were available, visually inspected the spectrograms of the voice samples and noted some effects of the workload-induced stress. Some subjects consistently displayed an increase in frequency and amplitude during the high stress tasks, but for other subjects, just the opposite occurred.

Subsèquent research has suggested that a rise in frequency and amplitude is the more common response to workload-induced stress (Williams & Stevens, 1972, 1981; Shipp, Brenner, & Dohery, 1986). However, there was a great deal of variability across subjects observed in these studies. For example, Shipp, Brenner, and Doherty (1986) required their subjects to perform a computerized manual task under two levels of difficulty while counting out loud. The fundamental frequency, loudness, and tempo of the counting increased as the difficulty of the manual task increased. However, there was substantial variability across subjects, especially at the high workload level, and the differences between the workload conditions escaped statistical significance.

Some studies have reported increases in the jitter in the voice with increased workload (Lieberman, 1963; Brenner, 1986). Jitter is a measure in the variance in the frequency, reflecting the extent to which the voice deviates from a perfect sine wave at the fundamental frequency. However, this result was not replicated by Shipp, Brenner, and Doherty (1986).

During the 1970's there was some interest in detecting stress by using a "psychological stress evaluator" (Brenner, Branscomb, & Schwartz, 1979; Schifflett & Loikith, 1980). This device was supposed to detect a "microtremor" in the voice which reflected stress, or, some said, lying. The existence and significance of the phenomenon was widely disputed and interest waned.

Other research has investigated fear-induced stress. As one might expect, fear-induced stress brings about increases in the tempo, amplitude and frequency of the voice (Williams & Stevens, 1969; Kuroda, Fujiwara, Okamura, & Utsuki, 1976; Brenner, 1986). While no study has directly compared the effects of fear-induced and workload-induced stress, the literature would suggest that fear-induced stress may bring about the more sizable changes in the voice.

Cannings et al (1979) speculated that the effects of stress on the voice were attributable to the neuromuscular response of the lanynx and diaphragm. The increase in muscle tension would increase pitch, tempo and volume. However, since individuals vary in their neuromuscular response to stress, individual voices vary in their response to stress. Cannings et al (1979) wrote that the laryngal and respiratory changes caused by stress would be most apparent in lengthy, connected speech. The short utterances common to a flight deck permit the speaker to breathe often while speaking, which might mask the effects of stress.

In sum, stress can cause an increase in the amplitude and frequency of the voice, as well as the rate of speech. Two major problems exist in applying these findings in actual operational environments in order to detect stressful situations. First, there is a great deal of interindividual variability in these effects; and secondly, there may be problems in obtaining voice samples that will reliably reveal these effects. Despite these limitations, researchers have begun to use devices to monitor the fundamental frequency of pilots' voices in order to detect stress (e.g., Peckham, 1979). It appears that with further research, reliable techniques for voice analysis will be able to reveal stress accurately in many individuals.

A U.S. court has recently admitted testimony concerning the analysis of a pilot's voice, recorded just before an aircraft disaster. In the case involved, it was crucial to know precisely what moment the pilot became aware that the aircraft was in difficulty. Brenner's deposition (Hoppie et al v. Cessna Aircraft Company, 1986) concerning the frequency and rate of the pilot's speech was used by the defendant in the case.

As voice analysis becomes more widely used, it must incorporate reliable and precise methodologies for assessing the workload experienced by the speaker. The purpose of the present study was to apply to workload research a voice analytic system that has already been extensively used in a clinical setting to

assess affect in psychopathological states. The utterances that were analyzed were short, imperative statements, similar in structure to those used on the advanced flight deck. The goal was to find the techniques for analyzing the voice of an individual to reveal the workload level.

METHODS

Subjects. Fifteen males, aged 18 to 47 (mean, 28.4; standard deviation, 7.4) participated. All reported normal hearing and normal or corrected vision. All were without physical handicaps, in good health, and spoke American English as his first language. Subjects received $8.00 per hour for participating, plus a $25.00 bonus for completing the tasks. Every subject did complete the tasks.

Procedure. Subjects were tested individually in a quiet, windowless room, 3.1 meters by 2.9 meters, free of decorations or other distractions. They sat on a comfortable, padded chair, facing a Taxan 630 computer monitor. This kind of monitor has especially rapid refresh and decay times, making it ideal for presenting moving figures. The monitor was connected to an IBM PC AT computer, but the keyboard was to the side so that it was out of view.

First, the principal investigator explained and demonstrated the tasks. The principal investigator answered any questions from the subject, and gave him the consent form to read and sign. All subjects signed the form.

Then, the subject rehearsed the tasks in two practice trials. The four actual trials followed.

Primary task. The voice samples were obtained using a task in which the subject was required to speak whenever one of two triangles on the monitor screen began to rotate. Two equilateral triangles, each 1.6 cm tall, appeared on the monitor screen. One was 8.3 cm to the right of the center of the screen, the other 8.3 cm to the left. At intervals ranging from 22.25 seconds to 22.75 seconds (mean, 22.5 seconds), one of the triangles, randomly chosen, began to rotate around its center point, at the rate of 55.4 degrees per second. The subject was required to say "Triangle please stop turning now" immediately when a triangle began to rotate. The subject was instructed to speak as he normally would, not less or more loudly or quickly. Fifteen voice samples were obtained in this way in each trial, except practice trials in which only 6 samples were obtained. Every subject was able to perform this task without omissions.

The subjects wore Sony DR-200 headphones with a head-held dynamic microphone attached. They heard 60 dB (0.0002 microbar reference) white noise through the headphones. This noise was to simulate ambient aircraft sounds. Speakers tend to be louder in the presence of noise (Alpert, 1966).

The voice was recorded on Maxell UR-60 cassette tapes, using an Onkyo TA-2058 cassette tape deck. The cassette tape deck was connected to a signal-activated switch, so that the switch closed when the subject was speaking. This signal-activated switch was wired to a Microsoft mouse so that one of the switches on the mouse closed whenever the subject spoke, and opened whenever the subject was silent. The computer was programmed to monitor the status of this switch on the mouse. In this way, the computer software could detect when the subject was speaking. The software recorded the time that elapsed between the start of the triangle rotation and the onset of the voice. The software stopped the triangle rotation and returned the triangle to its original position when the voice signal ended, or 6.5 seconds elapsed, whichever came first. In order to prevent momentary pauses from stopping the triangle rotation, the voice signal had to be off for 150 msec before the triangle stopped rotating.

Secondary task. The purpose of the secondary task was to vary workload. An increase in difficulty in the secondary task would increase cognitive loading, which might be reflected in changes in the acoustical measures of the voice samples obtained in the primary task.

The secondary task was a version of the Continuous Performance Task (CPT), which is used clinically as a measure of the ability to sustain vigilance (Rosvold, Mirsky, & Sarason, 1956). Numerals (white, 1.6 cm tall) were presented, one after the other, in the center of the computer screen. The numerals 1 through 6 were used. The subject was required to press a button, which he held in his hand, as quickly as possible whenever two successive numbers added to 7. For example, in the sequence "1 3 4 6 1 6 1" the subject was to press the button when the third, fifth, sixth, and seventh numerals appeared.

The software automatically recorded the number of omission, commission, late, and double strike errors. An omission error was scored when the subject failed to press the button when a target was present (i.e., the second number of a pair that added to 7). A commission error was scored when the subject pressed the button when no target was present. A late error was scored, instead of one commission error and one omission error, when the subject failed to press the button when a target was present but did press the button after the subsequent number, which was not a target. A double strike error was scored when the subject pressed the button twice, either correctly or in error, during a single number. The number of correct responses, defined as a single button press when a target was present, was also recorded.

Neither errors nor correct responses were recorded while the subject was speaking. It was felt that the act of speaking itself might alter the workload level.

The software calculated the error rate by dividing the total

number of omission, commission, and late errors by the total number of correct responses plus the total number of omission, commission, and late errors. Double strike errors were infrequent and were not included in the calculation.

Practice trials. The first two trials were practice trials, in which only six triangle rotations were presented. Since an average of 22.5 seconds elapsed between triangle rotations, each practice trial lasted less than 3 minutes. The first practice trial was a low workload trial. At the start of the trial, 800 msec elapsed between the numbers that were presented in the secondary task (the CPT). The software then continually adjusted this presentation rate to maintain the subject's error rate at .20. Every time forty numbers were presented, the software calculated the error rate during the presentation of those forty numbers. It then adjusted the rate of presentation by adding time to the internumber interval if the subject's error rate was greater than .20, subtracting time if the observed error rate was lower than .20. The amount of time to add or subtract was calculated by subtracting .20 from the observed error rate, multiplying the result by 6, rounding to the nearest integer, and multiplying by 50 msec.

The second practice trial was exactly like the first, except that it was a high workload trial. The initial interval between numbers was now only 450 msec, and the software attempted to maintain the error rate at .60.

By the end of each practice trial, the software had adjusted the presentation rate several times. The presentation rates that existed at the end of the two practice trials were respectively used as the initial presentation rates for the low and high workload experimental trials.

Experimental trials. After the practice trials, there was a ten minute break, during which time a research assistant orally administered the Borad test of lateral dominance (see appendix). Four experimental trials followed, with another ten minute break between the second and third trial. Two of the trials were high workload, two low workload. Half of the subjects received the trials in the order low-high-high-low, half in the order high-low-low-high.

Acoustical analyses. Each command ("Triangle please stop turning now") spoken by a subject in an experimental trial was called an "utterance" and subjected to analyses using the apparatus described earlier. Each utterance was composed of "peaks" which roughly correspond to syllables. As mentioned earlier, peaks are points of maximum amplitude relative to the values immediately preceding and following that point. The utterances collected in this study were all composed of between 5 and 7 peaks.

For each peak in each utterance, the following measures were

taken: amplitude, frequency, and duration. Amplitude was calculated in centibels relative to a 40 mv, 100 Hz square wave reference tone placed at the start of each cassette tape. An additional measure called "stress" was calculated by multiplying amplitude with frequency. Then, the means and variances for amplitude, frequency, duration, and stress were calculated for each utterance, using the data for the peaks that comprised the utterance. Then, the means of these variables, and the means of their variances, were calculated across all the utterances in the trial. These means were called the "grand means" for the trial.

Dropped subjects. One subject was dropped from the study because he was unable to achieve an error rate less than .4 in the low workload condition, no matter how slow the software set the CPT presentation rate. Another subject was run to take this subject's place. One subject was dropped from the acoustical analyses because of a technical problem with the cassette recorder.

RESULTS

Error rates. The results suggest that the software was successful in maintaining the error rates at approximately .6 and .2 for the high and low workload conditions respectively. The error rates for each trial are shown in Table 1. An analysis of variance was performed with two within-group factors: run (first versus second) and workload (low versus high). The main effect for workload was significant ($F [1,14] = 455.10$, $p<.0001$). However, there was also a main effect for run ($F [1,14] = 12.87$, $p<.003$), and a significant workload by run interaction ($F [1,14] = 14.82$, $p<.002$). These results suggest that the subjects' performance improved (i.e., their error rate decreased) between the first and second runs. Perhaps, this finding reflects the effect of practice. The significant interaction ocurred because the improvement was greater in the low workload condition than in the high workload condition (Newman-Keuls test, $p<.05$).

Internumber interval. As described earlier, the software continually adjusted the length of time that elapsed between the numbers presented in the Continuous Performance Test. When the subject's error rate was lower than the intended one (.2 or .6), the software sped up the presentation rate; when the error rate was higher than the intended rate, the software slowed the presentation. The mean internumber intervals that were actually required are shown in Table 2. Analysis of variance revealed a significant effect for workload ($F [1,14] = 144.13$, $p<.0001$). Of course, the presentation rate was faster in the high workload condition. There was also a significant effect for run ($F [1,14] = 46.07$, $p<.001$). This result once more suggests that the subjects' performance improved between the two runs. The software had to present the numbers more quickly in the second run in order to achieve similar error rates. The workload by run interaction was not significant.

CPT reaction time. The software calculated for each subject the mean and standard deviation of the reaction times for correct responses in the Continuous Performance Test. Table 3 shows the length of time in milliseconds that it took the subjects to press the button for correct responses on the CPT. Analysis of variance revealed only a main effect for workload ($F [1,14] = 338.76$, $p < .0001$). High workload brought about faster reaction times than low workload did. Run 1 did not differ significantly from run 2.

Table 4 shows the standard deviations of the reaction times, averaged across all subjects. Analysis of variance revealed a main effect for workload ($F [1,14] = 46.68$, $p < .0001$). There was also a main effect for run ($F [1,14] = 6.20$, $p < .026$). These results suggest that the high workload condition brought about faster reaction times, but the standard deviation of the reaction times was greater; there was more variability in reaction time, response by response. The main effect for run suggests that the variability in reaction time decreased in both workload conditions from run 1 to run 2.

Nature of errors committed. Analyses were performed to determine if the two workload conditions differed in the kinds of errors committed. The percentage of commission, omission, late, and double strike errors was calculated for each workload condition. Analysis of variance revealed two main effects for workload. First, high workload brought about a greater percentage (36.0 percent) of late errors than did low workload (23.1 percent) ($F [1,14] = 19.92$, $p < .001$). Second, high workload brought about a smaller percentage (2.7 percent) of double strike errors than did low workload (7.5 percent) ($F [1,14] = 11.62$, $p < .004$). These results stand to reason; the numbers were presented more quickly in the high workload condition, so the subjects had less time to press the button twice, and were more likely to press the button late.

Voice reaction time. The software recorded the number of milliseconds between the time at which a triangle began to rotate and the time at which the subject began to speak. This length of time was called "voice reaction time" and is shown in Table 5. Analysis of variance failed to reveal any significant effects for run or time, or for the interaction between the two.

Further analyses were to determine whether triangles on the left side brought about different voice reaction times than did triangles on the right. Half of the triangles that rotated were on the right, half on the left. Analyses of variance found no difference for the sides, and no interactions with the workload or run factors.

The analyses were repeated, but this time the triangles were classified as either being on the dominant side (e.g., on the right side for a right handed person) or on the non-dominant side. Four of the subjects were left handed on the handedness

scale. Analysis of variance yielded a significant interaction between side and run (see Table 6) (F [1,14] = 8.09, p<.013). There was a decrease in voice reaction time in the second run, as compared to the first run, only for triangles on the non-dominant side (Newman Keuls test, p<.05). There were no significant interactions with the workload factor.

Grand means of the acoustic measures. The first analyses of the acoustical measures were to compare their grand means. The design of the univariate analyses of variance was fully factorial, with two within-group factors: run (first run versus the second run), and workload (low versus high). The dependent measures were the grand means of the acoustical measures--that is, the means across all the utterances in a trial. The acoustical measures used were amplitude, frequency, duration, and stress, as well as the variances of those measures.

Figures 1 through 4 show the grand means for amplitude, frequency, stress, and duration. When the means of the two runs are calculated, high workload brought about higher frequency, amplitude, and stress than did the low workload condition. However, these trends did not reach significance. Frequency, amplitude, and stress tended to decrease from run 1 to run 2 but again, these trends did not reach significance.

The analyses of variance yielded no main effects for workload or run, and no interactions between the two factors, for any of these acoustical measures. There was therefore no evidence that the means or variances of the acoustical measures, when averaged across all the utterances in the trials, could differentiate the low and high workload conditions.

Individual differences. The effect of workload on the acoustical measures was not consistent across the subjects. Figures 5 through 8 show the distribution of the subjects with regard to the difference between the high and low workload conditions in amplitude, frequency, stress, and duration. The figures show that for each measure, there was a wide distribution. For example, the mean voice frequency for 6 of the 14 subjects was higher in the low workload condition than in the high workload condition. For the other 8 subjects, the reverse was true. Even though the present results and those of Shipp, Brenner, and Doherty (1986) suggest that high workload conditions might tend to bring a nonsignificant increase in the frequency of the voice, many individuals did not show this effect, or showed it only minimally. Similarly great individual differences occurred for the other measures.

Although there were these individual differences among the subjects, the subjects were, individually, quite consistent across the trials in their mean acoustical measures. Tables 7 through 10 show that the correlations among trials were quite high, especially for amplitude, frequency, and duration.

Temporal effects: amplitude. Further analyses were run to determine whether the acoustical measures changed in any systematic way over the course of the trials. These analyses went beyond simply analyzing the grand means of acoustical characteristics recorded over the span of a trial. These analyses were to detect systematic changes in the acoustical characteristics over time.

Figures 9 through 12 show the means across subjects for each utterance in each trial. Only 13 utterances are shown for each trial because several subjects spoiled some of their utterances with coughs. No subject spoiled more than 2 of the 15 utterances. All spoiled utterances were dropped form the analyses. In order to keep the number of utterances constant across subjects, only the first 13 utterances were used in the analyses when a subject correctly spoke more than 13.

Figures 9 through 12 reveal some trends across the course of the trials. For instance, the amplitude of the voice appears to diminish across each trial. In the low workload condition, the amplitudes of the utterances early in run 1 appear to be greater than the amplitudes of the utterances late in the trial. Then, at the start of run 2, the amplitudes were again at a relatively high level, which fell by the end of the trial. In the high workload condition, a similar drop in amplitude occurred during run 1. However, the amplitude did not return to a high level at the start of run 2. The amplitude at the start of run 2 was yet lower than the amplitude at the end of run 1. The amplitude fell during run 2 in the high workload condition, but it started at a relatively low level.

Statistical analyses supported the observation that amplitude did not remain constant over the course of the trials. Amplitude did vary across utterances in every trial (low workload, run 1, $F [12,156] = 1.81$, $p<.051$; low workload, run 2, $F [12,156] = 3.03$, $p<.001$; high workload, run 1, $F [12,156] = 1.75$, $p<.062$; high workload, run 2, $F [12,156] = 1.66$, $p<.081$).

Additional analyses were run to determine the statistical significance of the differences between workloads and between runs. The univariate analyses of variance used three within-subject factors: run (run 1 versus run 2), workload (low versus high), and utterance (first versus last). Only the first and last utterances were used in the analysis to remove the effects of fluctuations in the measures over the course of the trial. The comparisons between the first and last utterance were intended to show the overall direction of change during the trial.

In the analysis of the amplitude values, the run by workload by utterance interaction approached significance ($F [1,13] = 3.69$, $p<.077$). This interaction is graphed on Figure 13. This figure shows that the largest difference between workloads occurred during the first utterance in run 2 (although a posteriori tests

were not significant). This trend again suggests that in the low workload condition, amplitude fell across each trial, but started from a similarly high level in run 1 and run 2. In the high workload condition, however, the amplitude level fell during run 1 but never recovered to its earlier, higher level in run 2.

These apparent differences between runs 1 and 2 were further examined with a series of t tests which compared the two runs utterance by utterance. Separately for the low and high workloads, the first utterance in run 1 was compared with the first utterance in run 2, the second utterance in run 1 was compared with the second utterance in run 2, and so on for all 13 utterances in each run. The results revealed no differences between runs in the low workload condition. However, in the high workload condition, amplitudes in run 2 were lower than they were in run 1 for utterances 1, 7, 8, 9, and 10 ($t=2.70$, 2.20, 2.24, 2.68, 2.26 respectively, $p<.05$), as well as utterances 11 and 12 ($p<.10$). These results confirm the trends for amplitude that were apparent in the earlier analyses.

Temporal effects: frequency. Figure 10 shows the mean frequency of each utterance in each trial. The figure reveals that frequency remained relatively constant across utterances in the low workload condition. There appears to be a slight increase in frequency over the course of run 1. In the high workload condition, however, there appears to be a relatively large drop in frequency over the course of run 1.

Statistical analyses revealed no significant differences among the utterances in any trial except run 1 in the high workload condition ($F\ [12,156] = 2.46$, $p<.006$). A posteriori Newman Keuls tests ($p<.05$) revealed that this result largely reflected the high frequency present during the first utterance, relative to later utterances (the eighth and twelfth).

Univariate analysis of variance, using the design used for the amplitude measure, yielded a significant workload by utterance interaction ($F\ [1,13] = 7.92$, $p<.015$). This interaction is graphed in Figure 14. In the high workload condition, there was large drop-off in frequency during the first run, followed by a smaller drop-off in the second run, which started at a relatively low frequency. The dropoff in the high workload condition was significant ($p<.05$, Newman Keuls test).

T tests comparing the frequencies of the individual utterances in runs 1 and 2 failed to uncover any differences between the two runs in the low workload condition. However, in the high workload condition, there was a trend for the first several utterances to have lower frequencies in run 2, as compared to run 1. This trend reached significance for the fourth utterance ($t=2.43$, $p<.05$). This result is consistent with the trends revealed by the earlier analyses, that frequency tended to fall off more in the high workload condition than in the low workload condition, as the subjects said the sentence

repeatedly.

Temporal effects: Stress. Figure 11 shows the fluctuations across utterances in each trial for stress, which was calculated as the product of frequency and amplitude. Stress appears to have diminished across each run. Figure 15 shows the values of stress for the first and last utterances only. Stress fluctuated across utterances following a pattern similar to the one amplitude followed. Stress decreased during run 1 in both the high and low workload conditions. When run 2 began, however, stress returned to a high level in the low workload condition; no such recovery occurred in the high workload condition. Stress continued to fall from these respective levels during run 2. None of these effects was statistically significant, however.

T tests were again run to compare the stress measure of the utterances in runs 1 and 2. No t tests were significant in the low workload condition. However, in the high workload condition, utterances 1, 4 ($p<.10$), and 7 ($t=2.49$, $p<.05$) had lower stress measures in run 2 as compared to run 1. These findings again suggest that stress did return to its earlier high level at the start of run 2 in the low workload condition, but no recovery occurred in the high workload condition.

Temporal effects: duration. Figure 12 shows that there was a great deal of fluctuation over the course of the trials for the mean duration of the peaks. Figure 16 shows the data for the first and last utterances only. Statistical analyses of the data for the first and last utterances revealed an interaction between workload and utterance ($F [1,13] = 15.12$, $p<.002$). The duration of the peaks appears to have fallen over the course of the two low workload trials, but not during the high workload trials (Newman Keuls tests, $p<.05$). There may have been a general trend for the subjects to increase the speed of their speech over the course of the low workload trials. However, there was a great deal of fluctuation in peak duration over the course of the trials.

An attempt was made to replicate this finding using as the dependent variable, the total duration of the entire utterance. The results failed to yield any statistical significant findings.

Temporal effects: Variances of the measures. The preceding analyses concerned the means of the acoustic measures. The variances were also calculated and analyzed. Each utterance was comprised of 5 to 7 peaks. The frequency, amplitude, stress, and duration of each peak was determined, and then the mean and variance of each of these measures were calculated for each utterance. The variances have been used in other studies as measures in their own right. The variance of amplitude and frequency, for example, can reflect the amount of inflection, or of "flatness", in the voice. In the present study, however, no analysis concerning the variance of amplitude, frequency, or stress, yielded significant results.

Table 1
Error Rates

	Run 1	Run 2
Low workload	.273 (.094)	.218 (.062)
High workload	.619 (.031)	.610 (.020)

Note -- Entries are the observed error rates on the secondary task (the Continuous Performance Test). Standard deviations are in parentheses.

Table 2
CPT: Internumber interval (msec)

	Run 1	Run 2
Low workload	924.75 (156.95)	891.35 (151.20)
High workload	593.20 (85.05)	576.45 (76.20)

Note -- Entries are the mean internumber intervals (in msec) that the software had to use to bring about the intended error rates on the Continuous Performance Test. The intended error rate in the low workload condition was .2; in the high workload condition, .6. Standard deviations are in parentheses.

Table 3
CPT: Reaction Time

	Run 1	Run 2
Low workload	577 (88)	564 (81)
High workload	358 (74)	355 (74)

Note -- Entries are the mean reaction times in msec of button presses in correct responses on the Continuous Performance Test. Standard deviations are in parentheses.

Table 4
CPT: Reaction Time
Standard Deviation

	Run 1	Run 2
Low workload	155 (35)	142 (33)
High workload	203 (38)	189 (29)

Note -- Entries are the means (and standard deviations in parentheses) of the standard deviations of the reaction times on the Continuous Performance Test. These standard deviations, in msec, were calculated across all correct responses in a trial.

Table 5
Voice Reaction Time (msec)

	Run 1	Run 2
Low workload	179 (152)	177 (55)
High workload	219 (152)	183 (80)

Note -- Entries are the mean reaction times in the primary task; that is, the times in msec between the start of triangle rotation and the onset of the voice. Standard deviations are in parentheses.

Table 6
Voice Reaction Time
Dominant Versus Non-Dominant Side (msec)

	Run 1	Run 2
Dominant	154	179
Non-Dominant	230	148

Note -- Entries are the mean voice reaction times in msec. Triangles were located on the left and right sides. Triangles were classified as on the dominant or non-dominant side on the basis of a handedness scale administered to each subject.

Table 7
Correlation matrix: Grand mean of amplitude

	Run 1 Low workload	Run 1 High workload	Run 2 Low workload
Run 1 High workload	.81**		
Run 2 Low workload	.64*	.79**	
Run 2 High workload	.69*	.72*	.77**

Note -- Entries are Pearson product-moment correlations between the grand means across all utterances in the trials shown, for the fourteen subjects.

*$p<.01$ **$p<.001$

Table 8
Correlation matrix: Grand mean of frequency

	Run 1 Low workload	Run 1 High workload	Run 2 Low workload
Run 1 High workload	.71*		
Run 2 Low workload	.65*	.67*	
Run 2 High workload	.68*	.85**	.85**

*$p<.01$ **$p<.001$

Table 9
Correlation matrix: Grand mean of stress

	Run 1 Low workload	Run 1 High workload	Run 2 Low workload
Run 1 High workload	.80***		
Run 2 Low workload	.58*	.61*	
Run 2 High workload	.52	.73**	.66**

*$p<.05$ **$p<.01$ ***$p<.001$

Table 10
Correlation matrix: Grand mean of duration

	Run 1 Low workload	Run 1 High workload	Run 2 Low workload
Run 1 High workload	.89**		
Run 2 Low workload	.90**	.86**	
Run 2 High workload	.77**	.77**	.79**

*$p < .01$ **$p < .001$

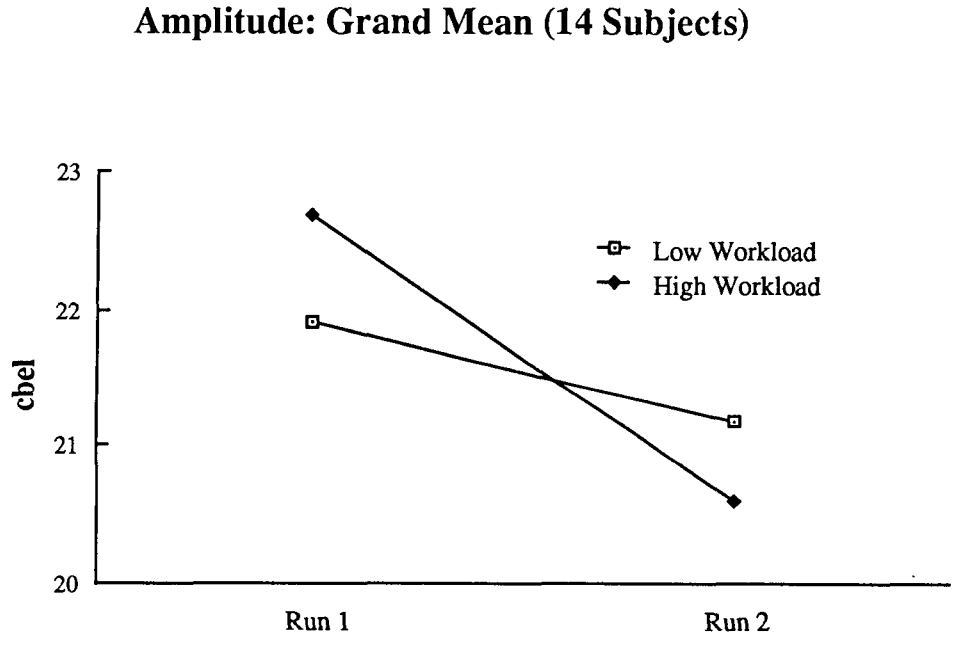


Fig. 1. The average amplitude of runs 1 and 2 was higher, though not significantly, in the high workload condition than it was in the low workload condition. This result was also obtained by Brenner (1986). Amplitude tended to fall from run 1 to run 2, but again the effect was not significant.

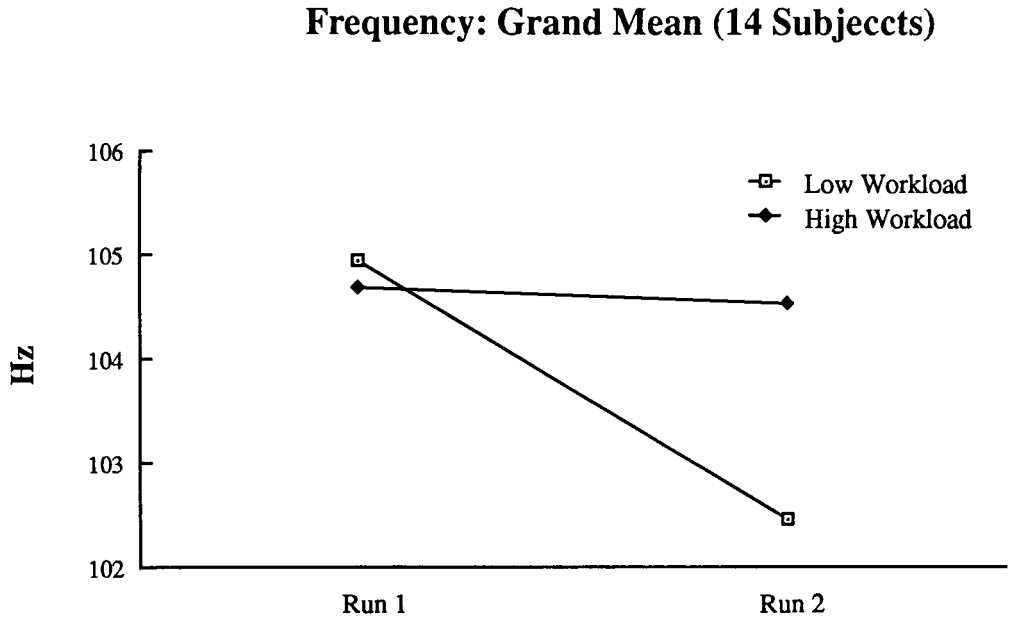


Fig. 2. The average frequency of runs 1 and 2 was higher, though not significantly, in the high workload condition than it was in the low workload condition. Again, this result was also obtained by Brenner (1986). Like amplitude, frequency tended to fall from run 1 to run 2, although this effect was not significant.

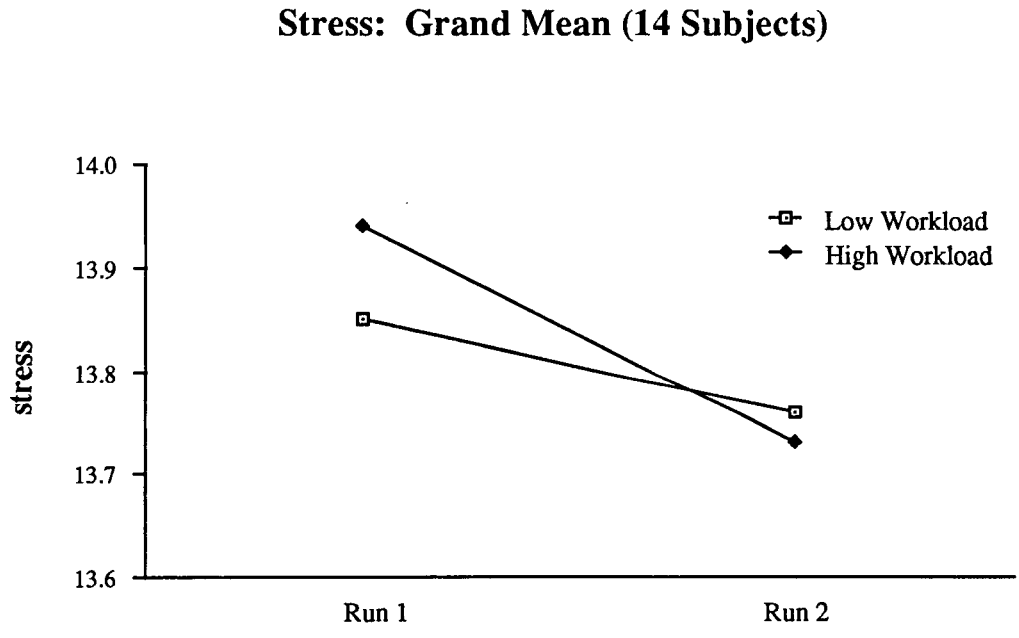


Fig. 3. Stress was the product of amplitude and frequency, and followed a pattern similar to both. It was higher, on average, in the high workload condition than in the low workload condition, and decreased from run 1 to run 2. Neither effect was significant. The decrease from run 1 to run 2 may be related to the subjects' improved performance on the Continuous PerformanceTest between runs.

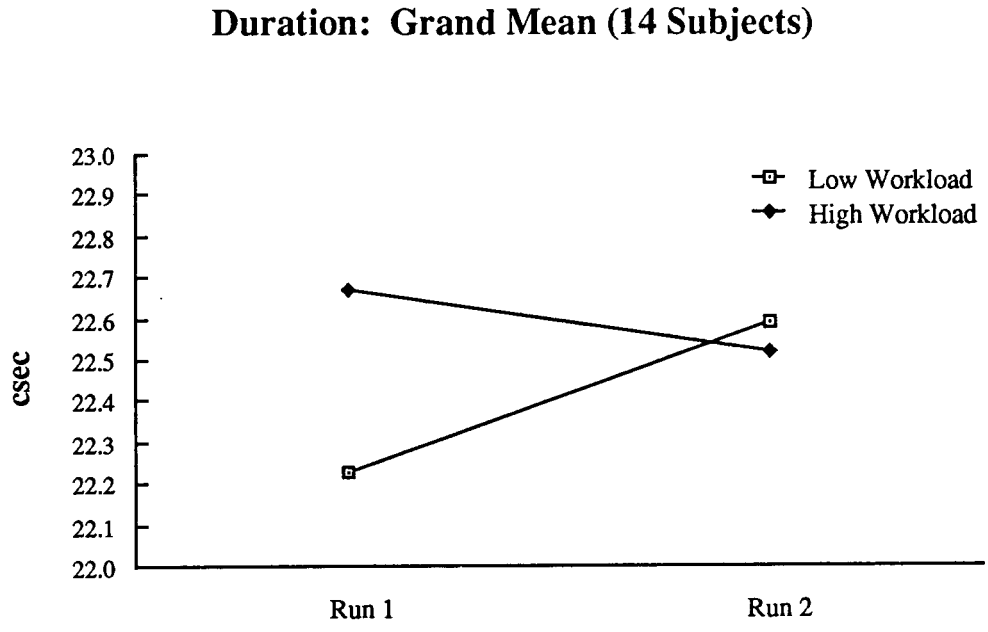


Fig. 4. No statistically significant findings emerged in the analysis of the grand means of the duration of the peaks.

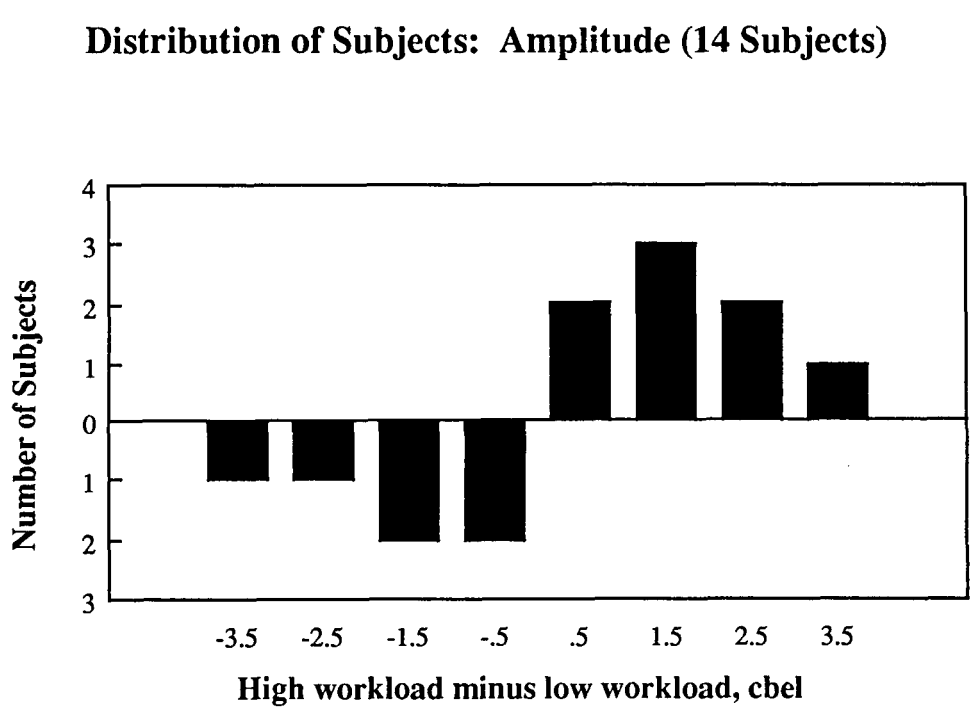


Fig. 5. There were large differences among subjects in how workload affected amplitude. For almost half of the subjects, high workload brought about lower amplitudes.

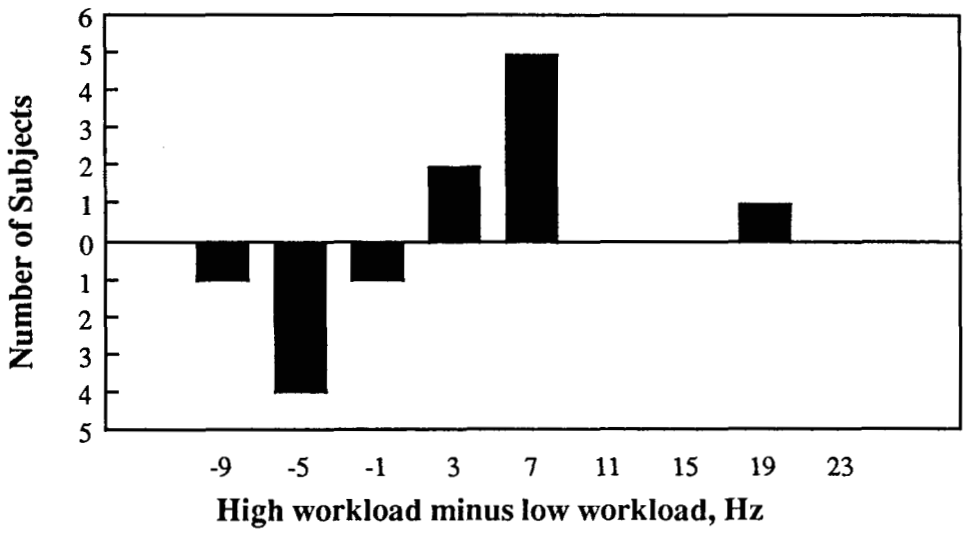


Fig. 6. There were also large differences among subjects in how workload affected frequency. Again, for almost half of the subjects, high workload brought about lower frequency.

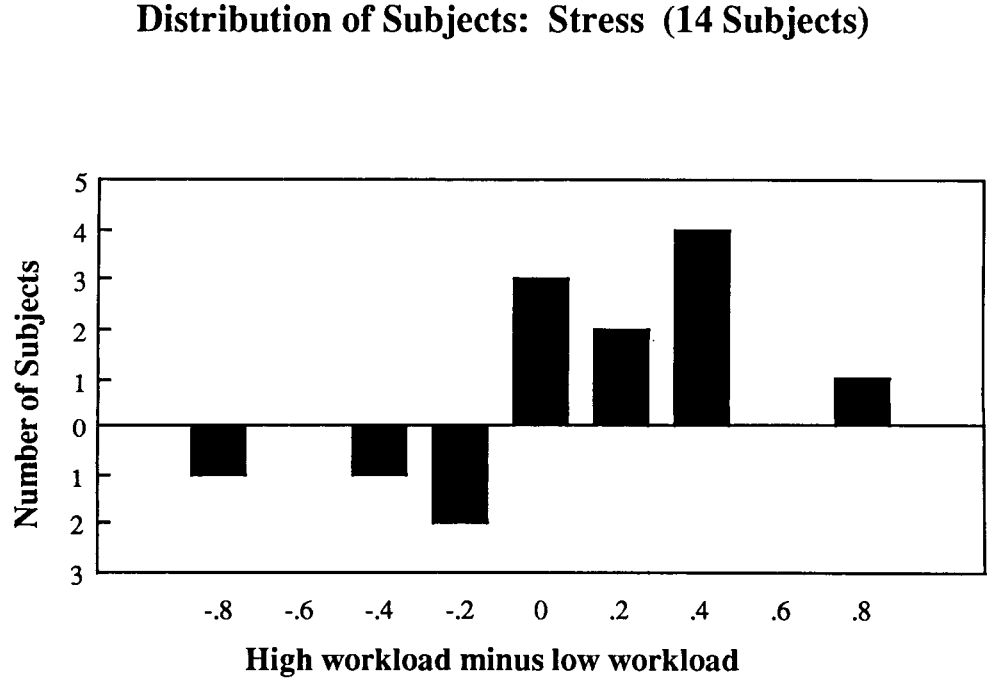


Fig. 7. For four of the fourteen subjects, high workload brought about lower stress measures.

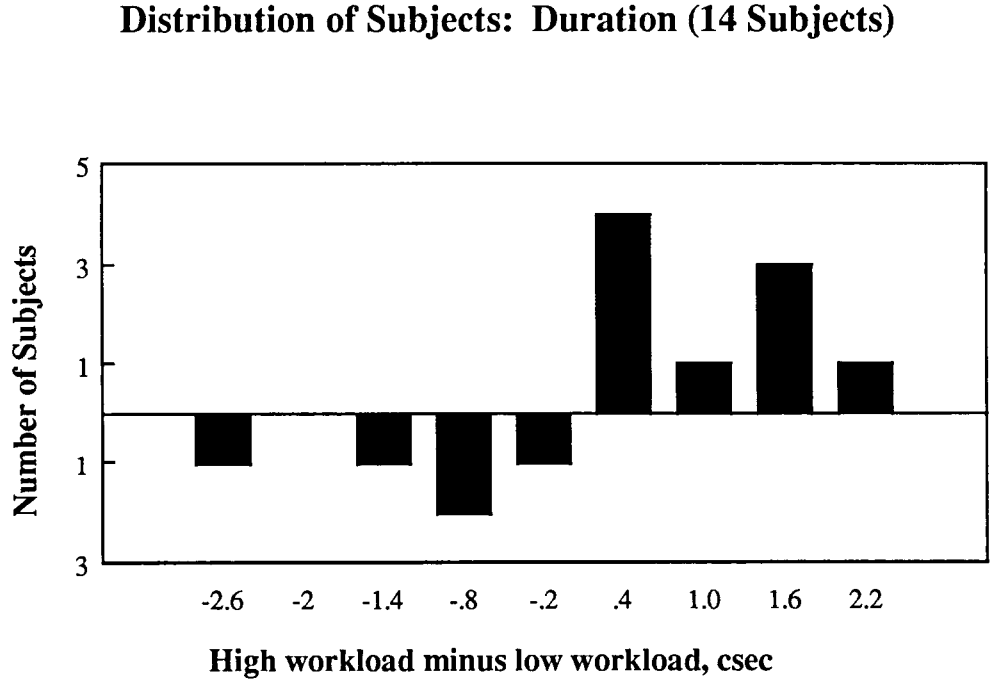


Fig. 8. There were large differences among subjects in how workload affected duration of the peaks.

MEAN AMPLITUDE (14 SUBJECTS)

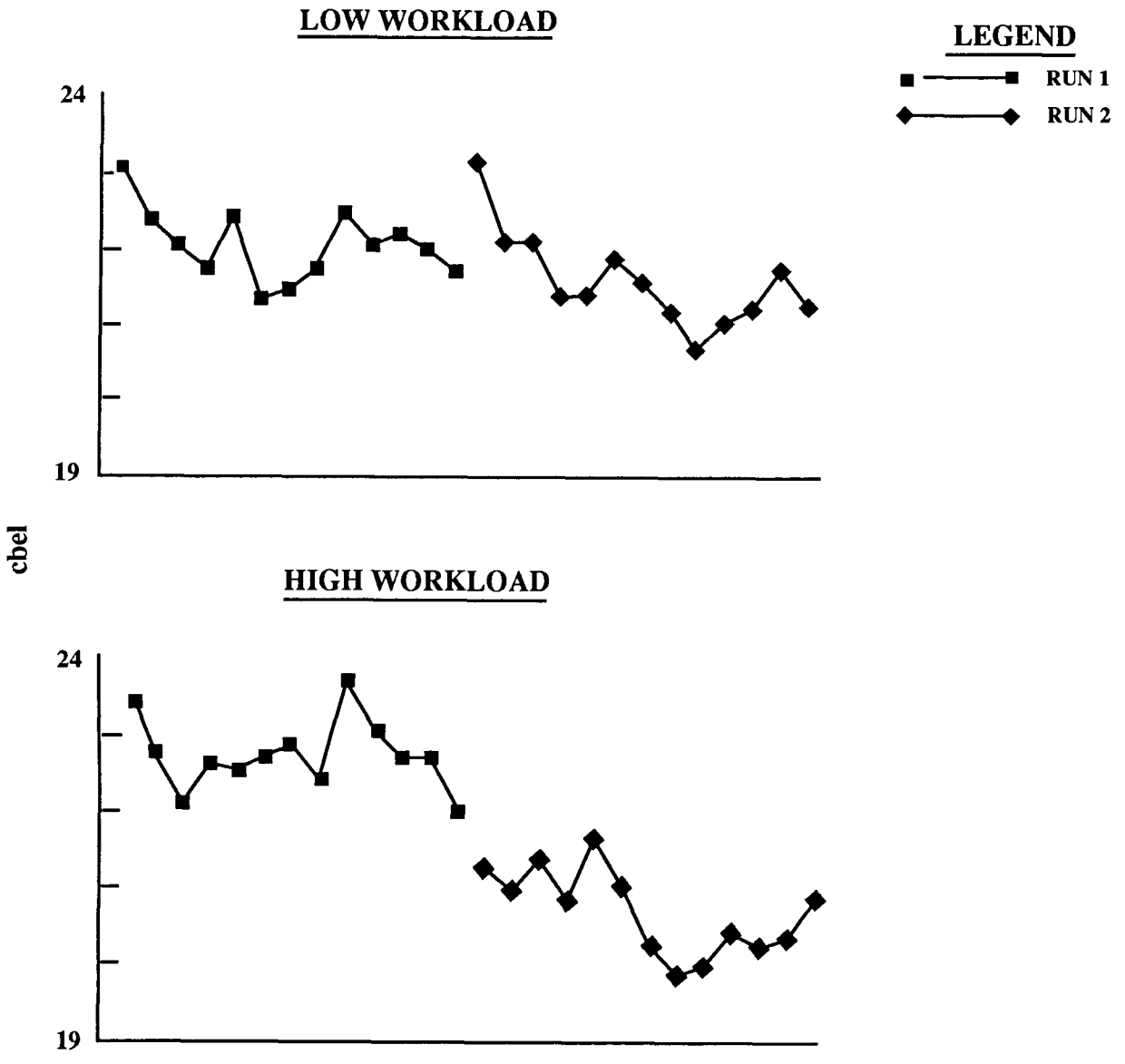


Fig. 9. Amplitudes fell over time in every trial. In the low workload condition, amplitude started from a similarly high level in runs 1 and 2. In the high workload condition, the amplitude never recovered to its earlier, higher level in run 2. T tests between utterances in the two runs demonstrated this effect.

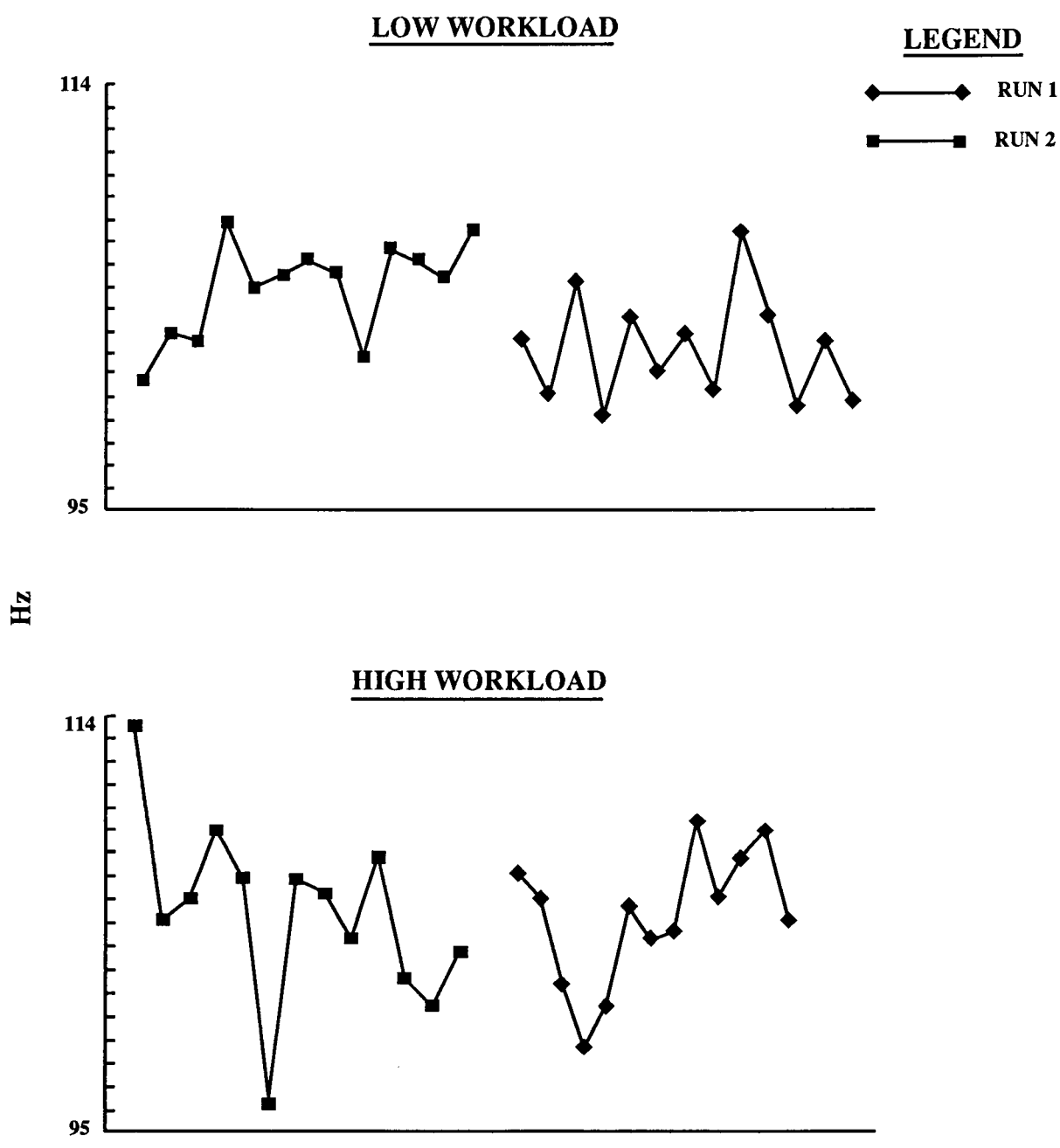


FIG. 10. Frequency fell in run 1 and the early part of run 2 in the high workload condition. This dropoff was statistically significant. In the low workload condition, frequency rose, though not significantly, during run 1.

MEAN STRESS (14 SUBJECTS)

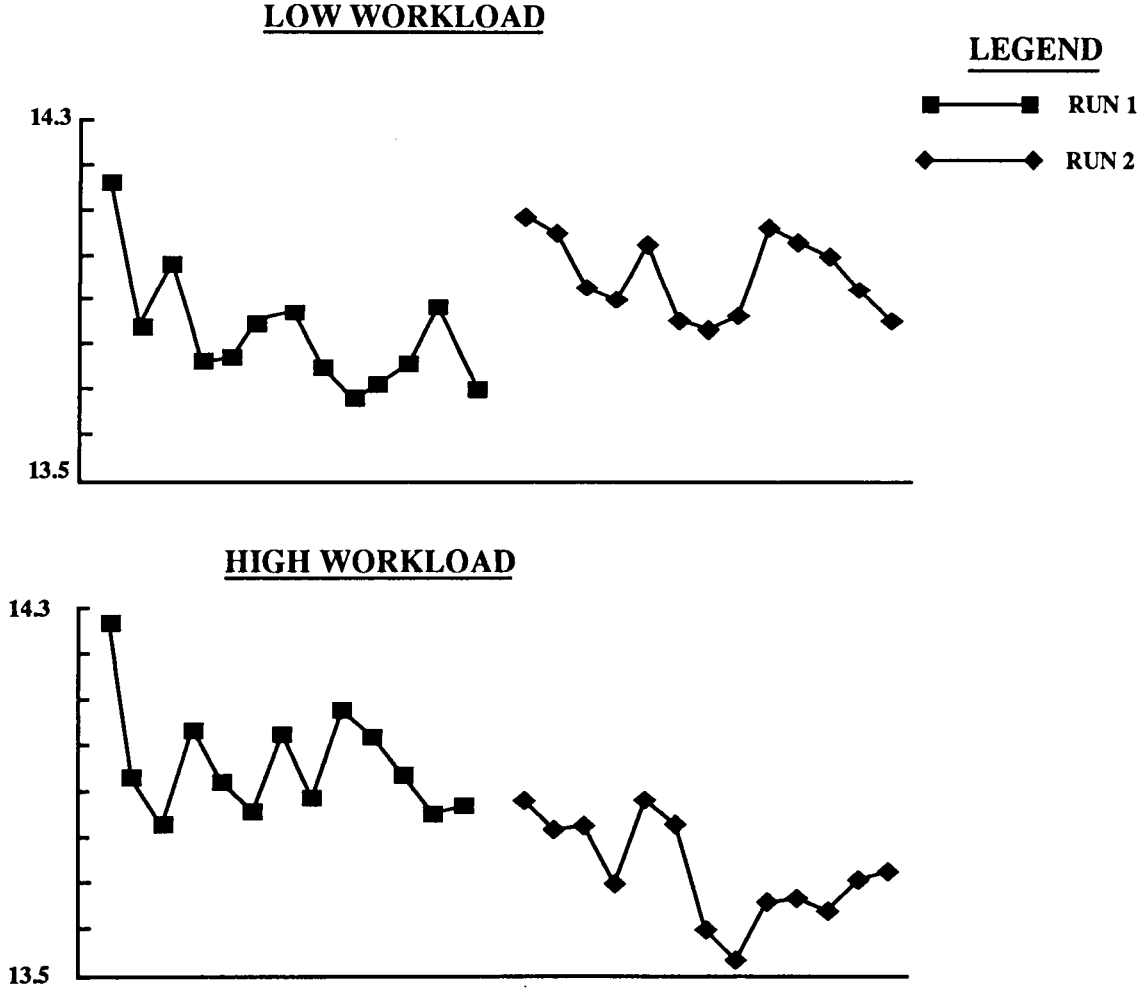


Fig. 11. Stress measures fell during run 1, but then returned to their earlier, higher levels at the start of run 2 in the low workload condition. No such recovery ocurred in the high workload condition.

MEAN DURATION OF PEAKS (14 SUBJECTS)

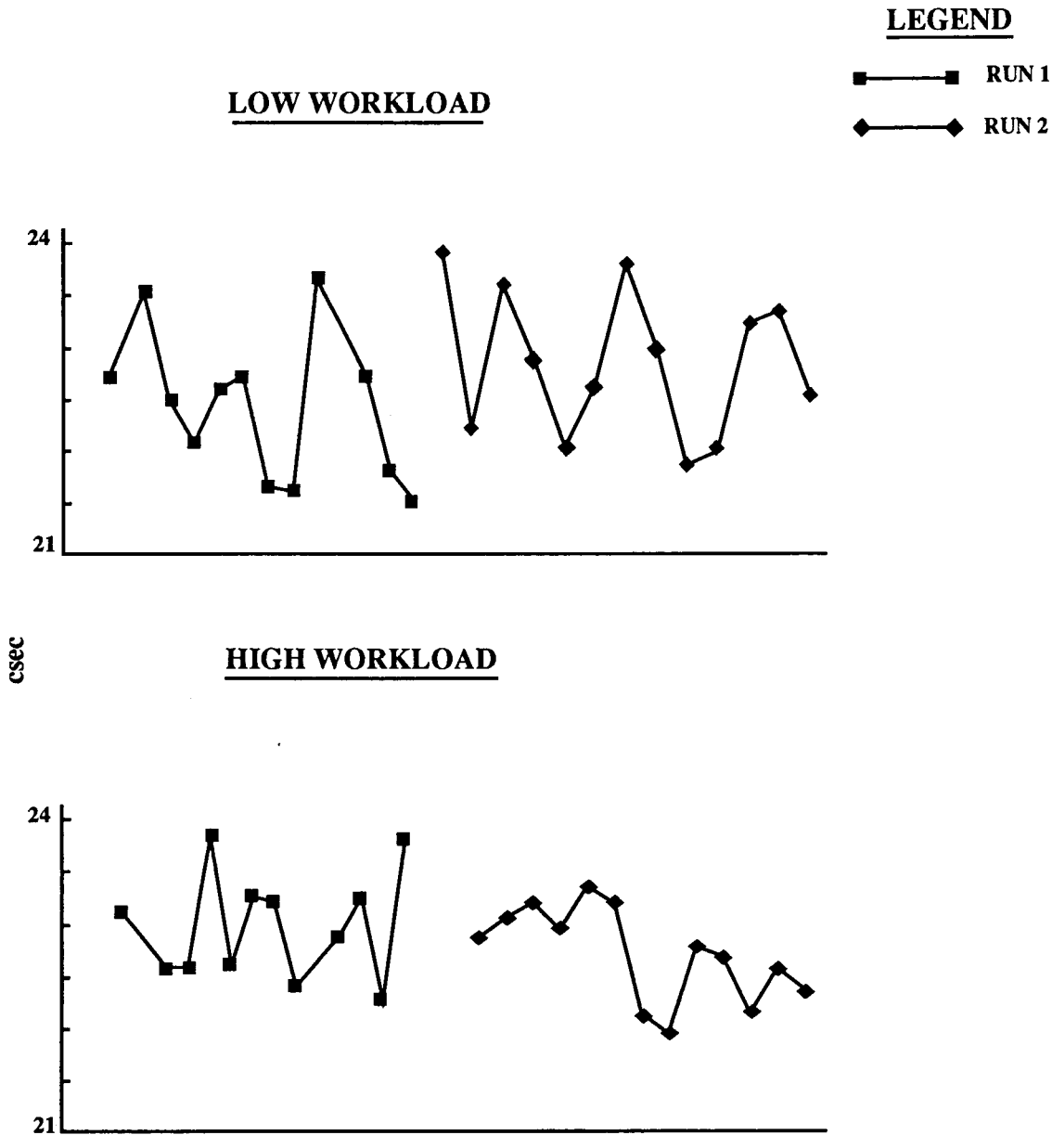


Fig. 12. Peak duration fluctuated a great deal across utterances in both workload conditions. There may have been a general trend for subjects to speak faster toward the late parts of the trials, particvlarly in the low workload condition.

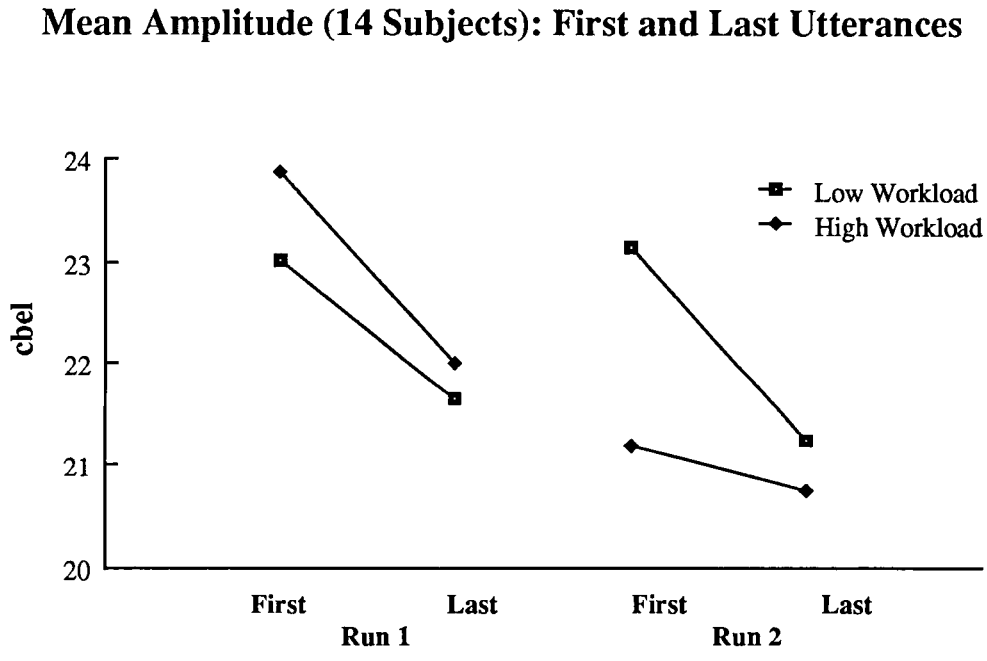


Fig. 13. The run by work by utterance interaction approached significance ($p < 0.077$). The first utterance in run 2 in the high workload condition failed to recover to earlier, higher amplitude levels.

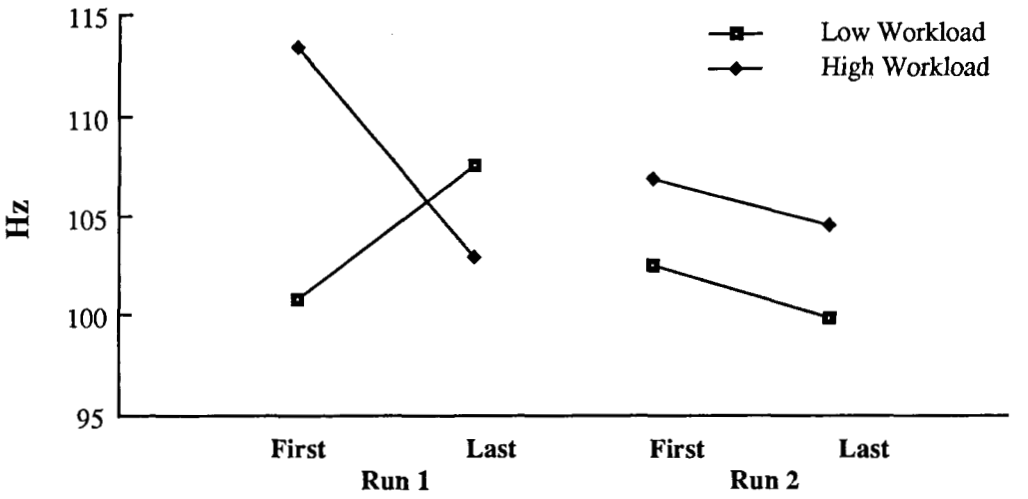


Fig. 14. The significant run by workload interaction was attributable to the dropoff in frequency in the high workload condition, particularly during run 1. The rise in frequency during run 1 in the low workload condition was not significant.

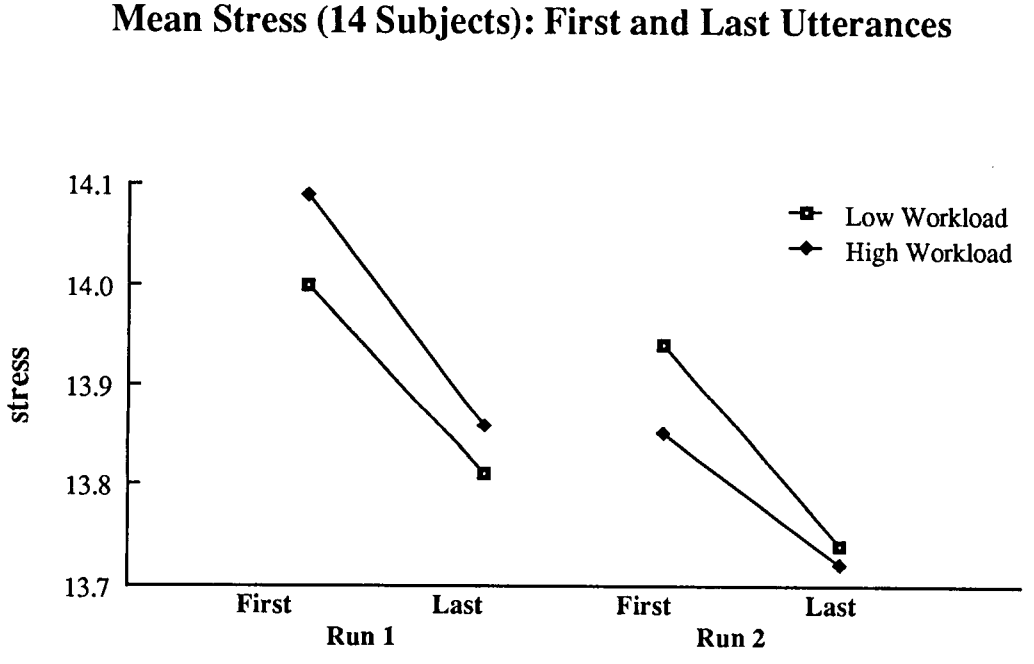


Fig. 15. Like amplitude and frequency, the stress measures fell during run 1 in the high workload condition, and never recovered to their earlier higher levels in run 2.

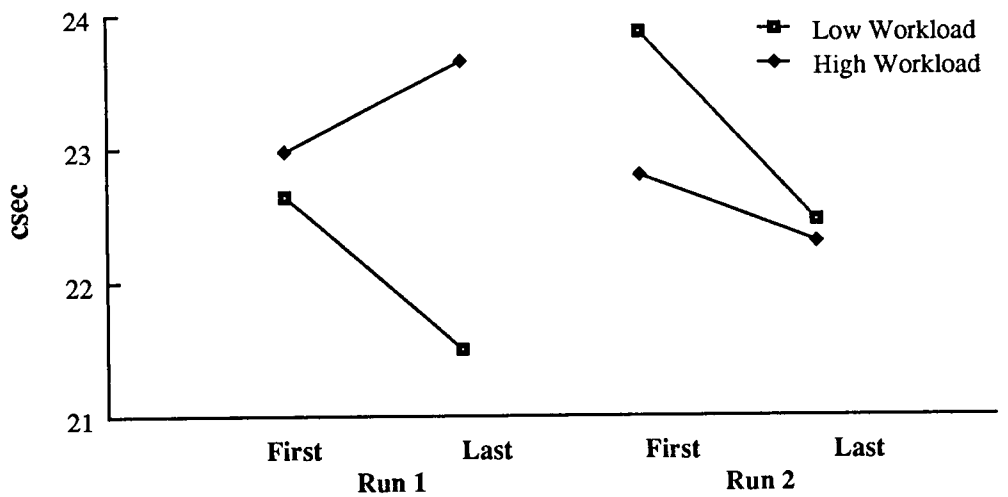


Fig. 16. There may have been a trend for the subjects to speak more quickly as the trials wore on in the low workload condition.

Peaks. Further analyses were run to determine whether workload had any systematic effect in the peak by peak articulation of the utterances. The univariate analyses of variance had three within-group factors: run (first versus second), workload (low versus high), and peak. Separate analyses were run for 5, 6, and 7 peak utterances. The dependent variables were frequency, amplitude, and duration. There were no statistically significant results in the analyses, other than the main effect for peak, which was consistently significant. The result suggests that the subjects had characteristic ways of articulating the sentence, and that these ways were relatively unaffected by workload.

DISCUSSION

The results suggest that the software for the Continuous Performance Test succeeded in changing the demands placed upon the subjects; the error rates during the high workload condition were significantly greater than the error rates during the low workload condition. By continually adjusting the rate at which the numbers were presented, the software held the error rate to approximately .6 during the high workload condition, and approximately .2 in the low workload condition. However, in the low workload condition, the subjects' error rates decreased from the first trial to the second. The reason for this effect may be that during the first trial, the subjects needed more practice before they could reach an error rate as low as .2. As a result, their mean error rate during the first low workload trial reached only .273. It was closer to .2, at .218, by the second low workload trial. Perhaps, this learning effect could have been removed by giving the subjects more practice before the trials or by making the low workload trial more difficult by requiring a higher error rate, such as .3.

The error rates could be used as a measure of the workload placed upon the subjects. In the present study, the error rates reflected a significantly higher level of workload during the high workload condition, as compared with the low workload condition. In future research, several levels of workload could be presented, each bringing about a different error rate.

In the present study, the subjects' mean reaction times in the Continuous Performance Test were faster in the high workload condition than they were in the low workload condition. At the same time, the standard deviations of the reaction times were greater. This result suggests that subjects were forced to respond more quickly in the high workload condition, and that their ability to do so fluctuated. These findings suggest that the high workload condition did put relatively greater demands upon the subjects.

At the time that this study was designed, it was felt that the two workload conditions would bring about differences in the mean acoustical measures observed during each trial. Accordingly, the mean for each acoustical measure was calculated for each subject, in each workload condition, in each run. No statistically significant effects emerged when the workload

conditions and the runs were compared. Frequency, amplitude, and stress were higher in the high workload condition, but these effects were not statistically significant. Shipp, Brenner, and Doherty (1986) reported a very similar finding. However, they also used a baseline condition in which the subject performed very little work. The baseline condition, compared with the workload conditions, brought about significantly lower frequency and amplitude.

The mean amplitude, frequency, and stress decreased, though not significantly, between run 1 and run 2. It is possible that these trends are related to the improvement in performance on the CPT between runs. There was thus some evidence that increased workload was related to increased mean frequency and mean amplitude, but the relationship never reached significance in these data.

Inspection of the data revealed that one possible reason that the differences in frequency and amplitude between workloads did not reach significance is that these acoustical measures fluctuated utterance by utterance.

Analyses revealed that this fluctuation in the acoustical measures across utterances may have been systematic, at least for the measures amplitude, frequency, and stress. These measures fell over the course of the first run in the high workload condition. Then, in the second run in the high workload condition, amplitude, frequency, and stress began at a level that was lower than their respective levels at the start of run 1. The measures each decreased during run 2, but at a slower rate than during run 1. The decrease in frequency in run 2 was followed by a small increase.

In the low workload condition, the acoustical measures followed a different pattern. Amplitude and stress fell during run 1, much like they did in the high workload condition. However, in the low workload condition, the values for amplitude and stress at the start of run 2 were as high as they were at the start of run 1. There was no reduction in amplitude and stress at the start of run 2, as there was in the high workload condition.

In the low workload condition, frequency remained relatively constant across utterances. There was a small increase during run 1, and a small decrease during run 2. The significant decrease in frequency apparent in the high workload condition did not occur.

In sum, in the high workload condition, amplitude and frequency (and their product, stress) all began at a relatively high level, fell over the course of run 1, and never recovered to their old level at the start of run 2. Amplitude and stress fell during run 2, but started from a lower level. Frequency also fell during run 2, but increased late in the trial. The low

workload condition did not bring about this loss of energy over time. These patterns may reflect the effect of fatigue or "strain" upon frequency and amplitude. As the subjects labored to perform well in the high workload task, energy was drained from their voices. Cannings et al (1979) proposed that workload-induced stress affects the voice via neuromuscular changes to the diaphragm and larynx. Conceivably, this increased tension over time could cause fatigue, and a loss of loudness and pitch.

In the present study and in Shipp, Brenner, and Doherty's (1986) work, mean frequency and amplitude were elevated, though not significantly, in the high workload condition. Perhaps, this finding reflects the greater effort that the subjects had to devote to the tasks. However, this greater effort may take its toll over time in a way analogous to physical effort. For example, if a person were to lift a heavy weight repeatedly, the person would gradually lift it more slowly, or not as high--that is, with less energy. After a rest period, the person might resume lifting the weight with less energy than he had at the start of the first series of lifts. That pattern would not occur if the person lifted a much lighter weight.

The utterances used in the present study were similar to the ones that will be commonly repeated in the advanced flight deck: short, imperative sentences. The present results might have looked very different if normal conversation were used as the voice samples, or if relatively long periods of counting were used (as in Shipp, Brenner, & Doherty, 1986). The repeated, short utterances that will be common in the advanced flight deck make it possible to measure patterns of change in the acoustical properties of the aircrew's voices over time. Future research should examine the patterns of frequency and amplitude over a large number of trials, with rest periods interspersed, to get a better idea of how sustained high workload affects voice measures.

By determining the change in voice characteristics over time, it may be possible to detect high workload situations in aircrew environments. The most valuable technique may be to record the pattern of change in frequency and amplitude over time, and especially to note the frequency and amplitude after rest periods. This technique would not have to take into account individual differences in the effect of workload on the overall means of acoustical measures. It would require only the detection of characteristic changes in voice measures over the course of the aircrew's tasks. A rapid drop in the energy of the voice, as reflected in amplitude and frequency, followed by the failure of the voice to achieve old energy levels after rest periods, can signal that the demands of the situation are taking a toll on the speaker. In this way, voice measures, alone, or in combination with psychophysiological measures, may provide one method for evaluating the demands placed upon aircrews.

VOICE ANALYSIS ON THE ADVANCED FLIGHT DECK

The voice analytic apparatus had been used, prior to the present study, primarily to analyze the speech of depressives and other psychiatric patients. There was little problem in applying the apparatus to the present study of workload. The voice analyses still occurred off-line, using tapes of the speakers' voices, and the statistical analyses of the acoustical data still occurred some time after the data were generated. However, this system would need to be modified in several ways in order to be used in the actual flight deck environment.

First, the system would need to be miniaturized, so that it fit on the flight deck. The speech filtering, processing and digitizing hardware would need to be packed onto a single circuit board. Existing, commercial voice recognition systems for the IBM PC typically have boards which incorporate signal processing hardware. However, it would be necessary to develop new boards, to work in a computer on the advanced flight deck, for the present application. Similar systems have been developed for the analysis of pilots' electrocardiograms during flight (Armstrong, 1985).

The computer itself would have to derive the acoustical data from the pilot's voice in real time, and immediately perform the statistical analyses necessary to detect high workload conditions. This real time capability would require some enhancements to the present software. However, no rewriting of the software would be needed to reduce execution time; the existing software is fast enough to be used in a real time application. Statistical procedures to detect drop-offs in voice amplitude and frequency would need to be written to operate along with the existing software.

The details of these statistical procedures must be determined in further research. One possible set of algorithms would first obtain samples of the pilot's voice during a no work, baseline period. The pilot would recite commands that might be issued on the flight deck, while the pilot was relaxed and not actually operating any equipment. The software would calculate from these baseline voice samples the mean and standard deviation of the amplitude and frequency of the pilot's voice. It would also calculate the rate at which the amplitude and frequency fell over the course of the baseline period.

The results of the present study suggested that workload affects the rate at which amplitude and frequency fall over time. The software on the on-board computer should be able to compare drop-offs in the amplitude and frequency in the pilot's voice during the flight to the drop-off observed during the baseline period. The magnitude of the drop-offs, and the length of time required to reach the lowest amplitude and frequency levels, would be calculated during the baseline period. Then, the length of time required for similar drop-offs would be observed during

the flight. The present study suggested that high workload conditions would bring about rapid drop-offs.

The software used by the voice analytic apparatus on the advanced flight deck should also be able to detect when the amplitude and frequency do not recover to earlier, higher levels following rest periods. The present study suggested that high workload is associated, at least in the laboratory, with this failure of the voice to recover. Further research may confirm that high workload tasks have a similar effect in the actual operating environment. If so, the software used for voice analysis should be able to detect periods of several minutes in which no voice commands are issued. The amplitude and frequency of the commands occurring immediately after these silent periods would be compared with the amplitude and frequency of the commands issued just before the silent period. The amplitude and frequency of the pilot's voice would be expected to fall over the course of a series of commands. If a silent period occurred after the series of commands, the software would determine what change occurred in the amplitude and frequency following this silent period. The amplitude and frequency might recover to earlier, higher levels, signaling a relatively low workload condition, or they might not recover, but continue falling to still lower levels, signaling a relatively high workload condition.

In sum, it should be possible to put voice analytic equipment on the advanced flight deck to assess pilot workload, while not interfering with the crew's activities. The voice analytic equipment used in the present study could be used on the advanced flight deck if it were miniaturized, and if the results obtained in the present study were applied in new software that performed the appropriate statistical analyses.

POTENTIAL LIMITATIONS

Any voice analytic system used on the advanced flight deck would have to overcome several potential problems, including: limitations on the number and kind of voice samples; aircraft noise; and discrepancies among individual voices in their responses to changes in workload.

Sampling constraints. The voice analytic apparatus used in this study had been used in previous studies to analyze both continuous, connected conversation, as well as isolated utterances. In the advanced flight deck, all the speech to be analyzed will be in the form of short, isolated utterances, like the ones subjects in the present study spoke. The voice analytic apparatus can quantify changes over time in the amplitude and frequency of speech in the advanced flight deck, much as it quantified those changes in the present laboratory study. If many devices on the advanced flight deck come under voice control, there will be a sufficient number of utterances for analyses to reveal the workload experienced by the speaker.

Noise. Voice samples in the present study were collected in a quiet laboratory room. In the operational environment, there will be ambient noise. The voice analytic procedures must be insensitive to this background noise so that it does not affect the acoustical measures.

The advanced flight deck will appear in civilian aircraft like the Boeing 757 and 767, where ambient noise levels in the cockpit are in the range of 60 to 76 dB (0.0002 microbar reference). The condenser-boom type of microphone that is commonly used in commercial aircraft allows the operator's voice to be plainly heard above this level of background noise. (White & Parks, 1985). The microphone is kept a constant distance from the mouth. The high signal-to-noise ratio should minimize the effect of noise upon the voice analytic procedure.

The voice recognition devices on the flight deck could work with a "hot mike," or microphone that is always on. However, "hot mikes" create the possibility of errors occurring through random voice inputs or transient noise. "Cold mikes" require a button press for the microphone to be active. Therefore, "cold mikes" partially defeat one purpose of the voice system, which is to eliminate manual actions. A compromise would have the microphone activated by a certain set of words, such as one of the verbs that start the commands in task oriented grammar. Such a system would be relatively unaffected by irrelevant speech and sounds.

While the design of the cockpit and the microphone system can minimize the unwanted noise that reaches the voice analytic system, it is still necessary to remove the effects of noise from the voice analysis. The equipment used in this study included a bandpass filter in order to remove much of the noise. The AC signal coming from the tapedeck was passed through this filter, which was set to restrict the signal to a range around the speaker's fundamental frequency. This filter eliminated some of the noise, as well as harmonic frequencies. The filter was adjusted for each particular speaker.

A second way in which noise is removed from the analysis is through the use of a threshold. The voice analysis software requires the researcher to input the amplitude of background noise that can be expected in the voice samples. The software does not consider the signal to be speech until its amplitude is a preset amount above the expected background level.

Despite all these precautions, some ambient noise will unavoidably contaminate the operator's voice that reaches the voice analytic apparatus. This noise will, to some small extent, boost the amplitude of the voice. Aircraft noise is likely to be relatively evenly distributed across frequencies, so the boost in amplitude is unlikely to shift the measure of the fundamental frequency. Because of the high signal to noise ratio, noise

should have minimal effect on the measures of the tempo of the speech.

The level of noise on the flight deck should be relatively constant across time, aside from predictable noise increases during take offs and landings. The present study suggested that drop-offs in voice amplitude and frequency over time can reflect workload level. The effect of noise will never be so great, or so variable, that an observed drop-off could be an artifact of a reduction in background noise. Therefore, the voice analytic apparatus used in this study should provide accurate data for workload assessment, even in the presence of aircraft noise.

Interindividual variability. Acoustical parameters of the voice have been found in many experiments to be affected by manipulations of stress or of workload. However, in every study, there were several subjects whose voices did not respond in the same way as the majority of the sample. For example, in one study, stress was induced by increasing the time pressure in a task (Hecker, Stevens, von Bismarck, & Williams, 1968). Some subjects were observed to consistently speak more softly when the stress was increased; others consistently spoke more loudly. Similar intersubject variability occurred for frequency. Such intersubject variability has been the rule, not the exception (Waskow, 1966; Williams & Stevens, 1972; Menahem, 1983).

Shipp, Brenner, and Doherty (1986) found that when the average values of amplitude and frequency were calculated for a group, increased workload brought about increases in amplitude and frequency. However, as workload increased, so did the variance of the amplitude and frequency; the voices of the subjects varied in their response to the increased workload. The average increases observed for the entire sample did not occur for every individual in the sample.

When voice analytic procedures are used on an actual flight deck, the goal is to detect changes in the workload experienced by one individual, such as the pilot. That one individual's response to workload may differ from the average response of any particular group. For example, some research has suggested that increased workload brings about higher frequency; however, there is a great deal of interindividual variability, so a pilot on the advanced flight deck might display no change, or lower frequency during an increase in workload.

The results of the present study suggest a solution to this problem. Instead of comparing the acoustical measures of an individual's voice with group norms, it is possible to compare the acoustical measures with prior measures from the same individual. Doing so takes advantage of the finding that the acoustical properties of an individual's voice are relatively stable over time. In the present study, correlations of each acoustic parameter across the experimental conditions were quite high. Changes that would be observed on the flight deck in the

acoustical parameters would probably not be attributable to random fluctuations; a more likely explanation would be changes in the mental state of the speaker.

The present study suggested that workload affects the rate at which the amplitude and frequency of the voice fall over time. Further research is needed to determine how this finding can be best applied in the advanced flight deck. However, the procedure suggested earlier would minimize the effects of interindividual variability. First, the drop-offs in amplitude and frequency would be observed in a no work, baseline condition. This profile of the operator's voice would serve as the comparison against which observed drop-offs in amplitude and frequency during flight could be compared. The present study suggested that high workload levels are revealed by rapid drop-offs in amplitude and frequency, and the failure of these measures to recover to higher levels following rest periods. Comparing drop-offs in amplitude and frequency this way avoids any reference to group norms and the problem of interindividual variability.

This procedure would immediately identify those individuals who do not display sizable drop-offs. They would be immediately apparent during the baseline period. Further research is needed to determine the percentage of individuals who do not display drop-offs. For these individuals, workload could not be assessed using the procedures outlined here. If amplitude and frequency drop-offs are almost universal, then the procedures could be used widely. It is possible that drop-offs are related to strain; high workload conditions bring on strain, and therefore reductions in the energy in the voice, faster than low workload conditions do. Since strain is common to all operators, it is possible that drop-offs are almost universal.

In sum, the problem of interindividual variability can be solved by confining statistical analyses to just the one individual under study. The rate at which the amplitude and frequency of the individual's voice fell during a test condition could be compared to that rate during a baseline condition. The rapidity of the drop-off may reflect the workload experienced by the speaker.

CONCLUSIONS

Previous research on workload and speech usually focussed upon the effect of workload on the mean values of acoustical measures like loudness, pitch, and tempo. While it has been possible to draw some general conclusions this way, the research has revealed substantial interindividual variability that has limited the sensitivity of that approach. The present study replicated some of this previous work (e.g., Shipp, Brenner, & Doherty, 1986) in showing that increased workload can bring general increases in mean amplitude and frequency, with a great deal of variability across subjects. The results of the present study, however, revealed a temporal trend in the effect of

workload that may be useful in assessing workload through its effect on speech.

In the present study, subjects spoke short, imperative sentences repeatedly, as pilots on the advanced flight deck would do. A simultaneous loading task was used to manipulate workload. There was a high and a low workload condition, presented in one of two orders: low-high-high-low and high-low-low-high. The results suggested that during both the high and the low workload conditions, the amplitude and, to a lesser extent, the frequency of the voice fell over the course of the tasks. This drop-off occurred more quickly in the high workload condition. In the low workload condition, the amplitude and frequency recovered to earlier, higher levels following rest periods. In the high workload condition, this recovery did not occur.

These results suggest that workload may not be best revealed by calculating the means of the acoustical parameters of the voice. Instead, it may be best to assess workload by the strain that it induces in the operator over time. Performing a high workload task over a period of time may cause strain, which in turn may reveal itself as a loss in the energy of the voice over time, that is, a reduction in pitch and loudness. Increased strain may also make it less likely that the voice would regain its energy following rest periods.

The effect of workload on voice, then, may be analogous to the effect of physical exertion on behavior. Everyone who repeatedly lifts a weight will eventually become fatigued and begin to struggle with the weight. This struggling will reveal itself in a loss of energy in the lifts, which will become slower or less numerous. It may take a long time before a strong person lifting a light weight begins to show this loss of energy. A weak person lifting a heavy weight would show this loss of energy more quickly. It may be possible to assess physical exertion by measuring how long it takes an individual who repeatedly lifts a standard weight to lose energy, and then comparing the length of time it takes the individual to lose energy with test weights. It is also possible to measure the extent to which the individual recovers energy after rest periods. These kinds of measures of physical exertion would use the individual's exertion with a reference weight as the standard against which exertion is assessed. Group norms and interindividual differences would not enter into the measures.

Analogously, it may be possible to measure how workload affects the voice of one individual. Applications in the advanced flight deck would allow investigators to detect how workload affected the voice of one particular operator, regardless of how that individual's response to workload differed from the responses of other people. It may be possible to measure how quickly the individual's voice loses energy, shown as a drop in loudness in pitch over time, while the individual is relaxed. After that, the workload imposed by a task could be

assessed by comparing how fast the individual's voice lost energy while he performed that task. The recovery of the energy of the voice to earlier, higher levels, following rest periods might also reveal the level of workload. Such measures of workload would be less obtrusive, and better tolerated by pilots than present measures which involve electrodes, like the electrocardiogram, or which involve questionnaires.

Further research is needed to determine whether voice measures of workload can be applied universally, or with only a limited number of individuals. Further research could also help refine the assessment technique as it moved out of the laboratory and into the actual operating environment of the advanced flight deck.

APPENDIX. Handedness Questionnaire

Please indicate for each of the items below left, both, or right.

Age_____

	Left	Both	Right
1. With which hand would you throw a ball to hit a target?	____	____	____
2. With which hand do you draw?	____	____	____
3. With which hand do you use an eraser on paper?	____	____	____
4. With which hand do you remove the top card when dealing?	____	____	____
5. With which foot do you kick a ball?	____	____	____
6. If you wanted to pick up a pebble with your toes, which foot would you use?	____	____	____
7. If you had to step up onto a chair, which foot would you place on the chair first?	____	____	____
8. Which foot would you use to stamp out a cigarette?	____	____	____
9. Which eye would you use to peep through a keyhole?	____	____	____
10. If you had to look into a dark bottle to see how full it was, which eye would you use?	____	____	____
11. Which eye would you use to sight down a rifle?	____	____	____
12. Which eye would you use to look through a telescope?	____	____	____
13. If you wanted to listen in on a conversation going on behind a closed door, which ear would you place against the door?	____	____	____
14. If you wanted to hear someone's heart beat, which ear would you place against their chest?	____	____	____
15. Into which ear would you place the earphone of a transistor radio?	____	____	____

References

Alpert, M. (1966) Feedback effects of audition and vocal effort on intensity of voice. Journal of the Acoustical Society of America, 39, 1218.

Alpert, M. (1982) Encoding of feelings in voice. In Clayton, P.J. & Barrett, J.E. (Eds.) Treatment of depression: Old controversies and new approaches. New York: Raven.

Alpert, M. & Anderson, L.T. (1977) Imagery mediation of vocal emphasis in flat affect. Archives of General Psychiatry, 124, 202-211.

Alpert, M., Homel, P., Merewether, F., Martz, J. & Lomask, M. (1986) Voxcom: A system for real time analysis of natural speech. Paper presented to the Eastern Psychological Association, New York City.

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

Armstrong, G.C. (1985) Computer-aided analysis of in-flight physiological measurement. Behavior Research Methods, Instruments, & Computers, 17, 183-185.

Armstrong, J.W., & Poock, G.K. (1981) Effect of operator mental loading on voice recognition system performance. Monterey, CA: Naval Postgraduate School Report NPS55-81-016, NTIS Access A107442.

Brenner, M. (1986) Analyzing tapes for psychological stress. Paper presented to the Air Law Symposium, Southern Methodist University.

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

Cannings, R., Borland, R.G., Hill, L.E., & Nicholson, A.N. (1979) Voice analysis and workload during the letdown, approach and landing. Paper presented to the Aerospace Medical Association, Washington, DC.

Easterbrook, J.A. (1959) The effect of emotion on cue utilization and the organization of behavior. Psychological Review, 66, 183-201.

Hart, S.G., Battiste, V., & Lester, P.T. (1984) POPCORN: A supervisory control simulation for workload and performance research. Paper presented to the Conference on Manual Control, Sunnyvale, California.

Hecker, M.H.L., Stevens, K.N., von Bismarck, G., & Williams, C.E. (1967) The effects of task-induced stress on speech. Contractor Final Report. Air Force Cambridge Research Laboratories report number AFCRL-67-0499, Office of Aerospace Research, USAF, Bedford, Massachusetts.

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

Hicks, T.G., & Wierwille, W.W. (1979) Comparison of five mental workload assessment procedures in a moving base driving simulator. Human Factors, 21, 129-143.

Hoppie et al v. Cessna Aircraft Company (1986) U.S. District Court, District of Montana-Missoula Division.

Kahneman, D. (1973) Attention and effort. Englewood Cliffs: Prentice-Hall.

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

Lea, W.A. (1983) Assessing speech recognizers for civilian airborne applications. Santa Barbara, California: Speech Science Publications.

Lieberman, P. (1963) Perturbations in vocal pitch. Journal of the Acoustical Society of America, 33, 597-603.

Mayer, M., Alpert, M., Stasny, P., et al. (1985) Multiple contributions to clinical presentation of flat affect in schizophrenia. Schizophrenia Bulletin, 11, 420-426.

Menahem, R. (1983) La voix et la communication des affects. Annee Psychologique, 83, 537-560.

New York Times (1987, August 20) Air crash inquiry focusing on flaps. Page B16.

North, R.A. & Bergeron, H. (1984) Systems concept for speech technology application in general aviation. In: American Institute of Aeronautics and Astronautics, Proceedings of the AIAA/IEEE sixth digital avionics systems conference, Baltimore MD, December 3-6, 1984. New York: American Institute of Aeronautics and Astronautics.

Ogden, G.D., Levine, J.M., & Eisner, E.J. (1979) Measurement of workload by secondary tasks. Human factors, 21, 529-548.

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. Technical report 79056, Royal Aircraft Establishment, Ministry of Defence, Farnborough, Hants, England.

Porubcansky, C.A. (1985) Speech technology: Present and future applications in the airborne environment. Aviation, Space, & Environmental Medicine, 56, 138-143.

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

Rolfe, J.M., & Lindsay, S.J. (1973) Flight deck environment and pilot workload: Biological measures of workload. Applied Ergonomics, 4, 199-206.

Rosvold, H.E., Mirsky, A.F., Sarason, I., et al. (1956) A continuous performance test of brain damage. Journal of Consulting Psychology, 20, 343-350.

Schiflett, S.G., & Loikith, B.S. (1980) Voice stress analysis as a measure of operator workload. Naval Air Test Center, Patuxent River, Maryland. Naval Report TM79-35Y.

Shipp, T., Brenner, M., & Doherty, E.T. (1986) Vocal indicators of psychophysiological stress induced by task loading. U.S. Air Force Report USAFSAM-TSQ-86-3.

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

Spettell, C.M. & Liebert, R.M. (1986) Training for safety in automated person-machine systems. American Psychologist, 41, 545-550.

Tole, J.R., Stephens, A.T., Vivaudou, M., Ephrath, A., & Young, L.R. (1983) Visual scanning behavior and pilot workload. NASA Contractor Report CR-3717.

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

Welford, R.W. (1978) Mental workload as a function of demand, capacity, and skill. Ergonomics, 21, 151-167.

White, R.W., & Parks, D.L. (1985) Study to determine potential flight applications and human factors design guidelines for voice recognition and synthesis systems. NASA Contractor Report CR-172590.

Wierwille, W.W. (1979) Physiological measure of aircrew mental

workload. Human Factors, 21, 575-593.

Wierwille, W.W., Guttman, J.C., Hicks, T.G., & Muto, W.H. (1977) Secondary task measurement of workload as a function of simulated vehicle dynamics and driving conditions. Human Factors, 19, 557-565.

Williams, C.E., & Stevens, K.N. (1969) On determining the emotional state of pilots during flight: An exploratory study. Aerospace Medicine, 40, 1369-1372.

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

Williams, C.E., & Stevens, K.N. (1981) Vocal correlates of emotional states. In J.K. Darby (Ed.) Speech evaluation in psychiatry. New York: Grune & Stratton.

Williges, R.C., & Wierville, W.W. (1979) Behavioral measures of aircrew mental workload. Human Factors, 21, 549-574.

Yntema, D.B., & Schulman, G.M. (1967) Response selection in keeping track of several things at once. Acta Psychologica, 27, 325-332.

NASA
National Aeronautics and Space Administration

Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA CR-4249		

4. Title and Subtitle	5. Report Date
Voice Measures of Workload in the Advanced Flight Deck	August 1989
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Sid J. Schneider, Murray Alpert, and Richard O'Donnell	
	10. Work Unit No. 505-67-11-01

9. Performing Organization Name and Address	11. Contract or Grant No.
Behavioral Health Systems, Inc. P.O. Box 547 Ossining, NY 10583	NAS1-18278
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address	Contractor Report
National Aeronautics and Space Administration Langley Research Center Hampton, VA 23665-5225	14. Sponsoring Agency Code

15. Supplementary Notes

Langley Technical Monitor: Randall L. Harris, Sr.

16. Abstract

Voice samples were obtained from 14 male subjects under high and low workload conditions. Acoustical analysis of the voice suggested that high workload conditions can be revealed by their effects on the voice over time. Aircrews in the advanced flight deck will be voicing short, imperative sentences repeatedly. A drop in the energy of the voice, as reflected by reductions in amplitude and frequency over time, and the failure to achieve old amplitude and frequency levels after rest periods, can signal that the workload demands of the situation are straining the speaker. This kind of measurement would be relatively unaffected by individual differences in acoustical measures.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Workload Voice Acoustical Analysis	Unclassified--Unlimited Subject category 54

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of pages	22. Price
Unclassified	Unclassified	72	A04

NASA-Langley, 1989

For sale by the National Technical Information Service, Springfield, Virginia 22161-2171