N88-23381
S11-53
140799
8P.
N0052345
BG535384

VOICE-STRESS MEASURE OF MENTAL WORKLOAD

Murray Alpert
New York University
New York, NY

Sid J. Schneider
Behavioral Health Systems, Inc.
Ossining, NY

In the 1970's, several studies employed voice analysis as a measure of workload. These studies usually looked at the suppression of the 8 to 12 Hertz microtremor in human voice as a measure of stress (Ref. 1). The existence and significance of the microtremor was controversial and the initial interest waned. Since then, a number of approaches have been developed, directed toward a detailed and extensive analysis of speech prosody. This method is intuitively appealing, since the emphasis, rhythm and inflection of a person's voice would seem to reflect psychological variables like stress, affect and the demands being placed on the speaker.

Research which explores the relationship between speech prosody and workload is relevant to the advanced flight deck. Flight crews will be making increased use of voice technology; the advanced flight deck will "speak" using voice synthesis and receive commands verbally from the crew. Therefore, speech samples will be readily available from flight crews in advanced flight decks. An apparatus that could assess the mental state of the flight crew from voice samples could be useful in the design and evaluation of the advanced flight deck.

The apparatus to be described was originally designed for applications in psychiatry, to provide objective and quantitative measures of variations in feeling states such as depression, mania, or the flat effect of schizophrenia (Ref. 2). Also of interest was the measurement of medication effects on such psychiatric states. Empirical studies have targeted the most discriminating acoustic parameters for each of these variables. Similarly, empirical studies could identify those acoustic measures which are most sensitive to the effects of variations in workload in aircraft situations.

This hybrid analog/digital analyzer provides information about such basic speech variables as fundamental frequency (pitch), amplitude (loudness), the duration of utterances and pauses, and the variances of these measures. The system consists of three main components: 1) a good quality stereo cassette tape deck; 2) a microcomputer equipped with an analog to digital conversion system (Northstar Horizon with a Tecmar TM-AD212 analog to digital converter); and 3) a multifunction analog signal processing unit. The analog computer unit provides the circuitry for filtering and transforming the speech signals prior to digital analysis. The raw AC signal that comes from the tape deck is first passed through a bandpass filter in order to restrict the signal to a range around the speaker's fundamental frequency, and to eliminate harmonic frequencies. The range between the filter can be adjusted for the particular voice; for example, it is usually set between 80 to 100 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, then full wave rectified to a DC signal and finally demodulated to produce a smooth signal that goes to another channel on the A/D converter board.

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 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.

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

1) Number of utterance, gaps, and peaks.

2) Mean and variance of the time durations of utterances, gaps, and peaks.

3) Mean and variance of the natural log of the amplitude of peaks as well as the log of the frequencies corresponding to those peaks.

4) The correlation between peak amplitude and peak frequency.

5) The distribution of peaks within utterances (i.e., how many 1 peak utterances, 2 peak utterances, etc. were there) as well as summary information about the duration of the peaks in those utterances.

The hardware and software allow for the setting of a threshold to eliminate background noise. It is also 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 analog 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 is uncontaminated by other speakers and noise that may be present in an aircrew operational setting. A calibration signal of known amplitude and frequency is recorded on the subject's channel. Since the subject uses a head mounted microphone of known output, the use of a calibration signal permits a usable estimate of the absolute voice level.

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 have suggested 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 (Ref. 3). 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 disturbance tends to be shown by long pauses and brief utterances (Ref. 4). Measurement of flat affect using this apparatus compared favorably to clinical ratings made by highly skilled attending psychiatrists in evaluating and predicting patient behaviors (Ref. 5).

Acoustic analysis has permitted the articulation of processes that are frequently confounded clinically and conceptually. Thus, it has become possible to distinguish effects from moods. Affects are encoded in voice emphasis. Affects are visible as the rapid fluctuations in the acoustic parameters amplitude and number of multi-peak utterances. Affects, such as excited affect, reflect momentary feelings of which the speaker may not be entirely conscious. Moods, on the other hand, are encoded in temporal patterns of utterances and pauses and have much slower temporal phases. Moods are the subjective feelings, like sadness and joy. They are revealed in the length 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. Changes in respiration will affect speech energetics; changes in muscle tone will alter the overtone structure and the speaker's voice quality (Ref. 6). These changes are visible as alterations in voice frequency lasting several 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, such as having the subject lie or by 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 controls. The different feeling states appear to be controlled by different and perhaps orthogonal brain mechanisms.

The apparatus has not 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. We have begun a study to determine which features of voice prosody reflect the workload experienced by the speaker. As of this writing, only two preliminary subjects have been run, but their results can be reported.

SUMMARY OF PROCEDURE

Subjects will be males between 18 and 50 years of age without uncorrected sight, speech, or hearing condition. Subjects will be run individually in a windowless room free of distractions. The subject will be seated at a Taxan 630 computer screen and have before him a hand held momentary contact switch. He will wear a set of headphones attached to a Shure head mounted microphone. White noise at 60 dB (0.0002 microbar reference) will be presented over the headphones to simulate cockpit noise.

An IBM PC AT computer will present simultaneous primary and secondary tasks. The primary task will be designed to be simple enough to be performed errorlessly. It will require speech and will be the source of the voice samples used for analysis. The secondary task will be used to manipulate workload. The system will continually monitor the error rate in the secondary task and adjust the presentation rate in order to keep the error rate constant. There will be a high workload (i.e., high error rate) and low workload (i.e., low error rate) condition.

In the secondary task, the numerals 1 through 6 will be presented in a predetermined order, one after the other, in the center of the screen. The subject's task is to press the button immediately whenever two numbers in a row total seven. Thirty percent of the numerals will be targets. While these numbers are presented, there are two triangles, one on either side of the central numbers, about 7 cm away. At intervals ranging from 18 to 28 sec, one or the other triangles, randomly chosen, will appear to rotate. The subject must state, as rapidly as possible after the triangle begins to rotate, "The triangle that started moving should stop now." The triangle will in fact stop rotating upon voice offset (or after 10 seconds pass). This speaking task is the primary task.

The computer will automatically record the number of correct responses in the secondary (number) task, as well as the reaction times of the correct responses. The number of commission, omission, double strike and late errors will also be recorded. In order to minimize the effect of speaking itself, the system will not record performance on the secondary (number) task while the subject is speaking as part of the primary task.

The reaction time of the voice in the primary task will be recorded. The voice itself will be captured on a cassette deck for analysis on the voice prosody analytic apparatus at the Millhauser Laboratories, New York University School of Medicine.

The session for each subject will begin with a series of short practice trials designed to familiarize the subject with the tasks. The trials will also suggest which presentation rates for the central number task result in error rates of 20 and 60 percent. These rates will be the initial presentation rates used, in the low and high workload conditions, respectively. The system is programmed to adjust the presentation rate at predetermined intervals to maintain the error rate at the desired level.

The low and high workload tasks will be presented in 10 minute segments. For half of the subjects, the order will be LHHL, and for the other half, HLLH.

Analyses of the cassettes should reveal which aspects of voice prosody are associated with increased workloads. The error rates and reaction times in the central number task will corroborate the assertion that mental loading has in fact been manipulated by changing the rate of number presentation. A faster presentation rate should increase error rate and decrease reaction time. The continual adjustment of the presentation rate will insure that workload transients are avoided to the extent possible.

Preliminary Results

Two preliminary subjects have been run in this study. Both were female. Their data will not be used in the final report of this research. Table 1 shows the error rates for the secondary (number) task in each of the two 10 minute runs in the high and low workload conditions. The table reveals that the software was effective in maintaining error rates close to the intended error rates. Table 2 shows the average reaction times for the button pushes in the correct responses in the number task. The high workload condition, in which the presentation of the numbers was fast, brought about faster reaction times than the low workload condition did. Table 3 shows that the voice reaction times (the length of time between the moment that the triangle started turning and the subject spoke) also tended to be faster in the high workload condition. The standard deviations of the voice and number tack reaction times, shown in parentheses in the tables, tended to be higher in the high workload condition, as compared with the low workload condition. These results would suggest that the speeded presentation of the numbers in the high workload condition was in fact successful in bringing about increased workload.

Table 4 shows the results of the analysis of the voice of the two subjects. The table reveals a trend for the frequency and amplitude of the voice to increase with each successive run, regardless of whether the run was in the high or low workload condition. However, these preliminary results suggest a possible trend for the voice frequency and amplitude to be higher, and the variance of the voice frequency to be lower, in the high workload condition. These results would replicate a previous study (ref. 6). However, this study must be run with the sample of 15 subjects before any conclusions can be drawn. Also, in this preliminary study, 25 voice samples were obtained in each run. This number of samples made the runs rather long and aversive to the subjects. In the actual study, the runs will be shortened to 15 samples. The means for the first 15 samples from the preliminary subjects were similar to the means for all 25 samples. By running a larger number of subjects in shorter runs, the effect of workload upon voice prosody should become more apparent.

REFERENCES

1. Schiflett, S.G. & Loikith, B.S.: Voice stress analysis as a measure of operator workload. Naval Air Test Center, Patuxent River, Maryland, 1980.

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

3. Mayer, M., Alpert, M., Stastny, P., et al.: Multiple contributions to clinical presentation of flat affect in schizophrenia. Schizophrenia Bulletin, vol. 11, 1985, pp.420-426.

4. Alpert, M.: Feedback effects of audition and vocal effort on intensity of voice. Journal of the Acoustical Society of America, vol. 39, 1966, pp. 1218.

5. Alpert, M. & Anderson, L.T.: Imagery mediation of vocal emphasis in flat

affect. Archives of General Psychiatry, vol. 124, 1985, pp. 202-211.

6. Shipp, T., Brenner, M., & Doherty, E. T.: Vocal indicators of psychophysiological stress induced by task loading. USAF SAM-TSQ-86-3, September, 1986.

ERROR RATES

	High Workload		Low Workload	
	First Run	Second Run	First Run	Second Run
DESIRED	.60	.60	.20	.20
SUBJECT 1	.64	.61	.19	.20
SUBJECT 2	.60	.62	.21	.20

Table 1. The desired error rate in the secondary (number) task in the high workload condition was .6; in the low workload condition, the desired error rate was .2. Error rate was defined as (number of errors) / (number of errors + number of correct responses). The software was able to maintain subjects' performance near the desired error rates.

NUMBER TASK REACTION TIME (msec)

	High Workload		Low Workload	
	First Run	Second Run	First Run	Second Run
SUBJECT 1	312 (204)	335 (194)	586 (126)	538 (127)
SUBJECT 2	329 (230)	353 (210)	663 (157)	652 (164)

Table 2. Entries are the mean reaction times in msec for correct responses in the secondary (number) task. The reaction time was defined as the time between the appearance of a target number (the second number of a pair that added to 7) and the moment the subject pressed the switch. Standard deviations are in parentheses. The high workload condition appears to have brought about faster reaction times and higher standard deviations.

VOICE REACTION TIME (msec)

	High Workload		Low Workload	
	First Run	Second Run	First Run	Second Run
SUBJECT 1	804 (104)	716 (114)	820 (79)	808 (83)
SUBJECT 2	988 (400)	1030 (263)	998 (213)	950 (191)

Table 3. Entries are the mean reaction times in msec for the primary (voice) task. The reaction time was defined as the time between the initiation of triangle movement and speech onset. Standard deviations are in parentheses. The high workload condition may have brought about faster reaction times and higher standard deviations.

Acoustic Measures

(summary data - 25 sentences per run)

	Work-load	Run #	Uttdur	Uttvar	Mean Amp	Var Amp	Mean Freq	Var Freq
SUBJECT 1	low	1	184.5	14.5	458.3	249.7	520.4	42.9
	high	2	185.2	2392.6	469.8	223.6	523.9	28.5
	high	3	211.6	5925.7	493.6	164.2	531.3	10.3
	low	4	205.2	3662.0	502.5	164.7	530.0	10.5
SUBJECT 2								
	high	1	180.4	3954.3	435.7	149.3	521.8	23.0
	low	2	180.4	1374.1	439.6	136.7	524.1	6.3
	low	3	190.6	126.3	437.1	202.3	521.5	7.4
	high	4	207.1	5336.6	448.4	169.7	528.4	5.6