VOICE STRESS ANALYSIS

Malcolm Brenner
National Transportation Safety Board
Washington, D. C.

and

Thomas Shipp
Speech Research Laboratory
Veterans Administration Medical Center
San Francisco, California

A voice measure of the speaker's physiological state has unique applications in the aerospace environment. Unlike other physiological measures, a voice measure is unobtrusive and does not require attaching any equipment to the person being tested. It can be employed in cockpit and spacecraft settings without interfering with ongoing activity and, if used on radio-transmitted speech, might be employed without any additional equipment in the flight environment. A voice measure can also be used on recorded speech as, for example, in accident investigation to determine the relative stress levels of different statements by the flightcrew for information relevant to human performance issues in the investigation. For the purposes of this paper, the term "stress" is used to mean changes in physiological state that result from changes in workload demands.

The aerospace community has been active in research on voice stress analysis (refs. 1 and 2). Although several aspects of the voice have been defined that appear to respond to psychological stress, it remains unclear from the research literature whether such voice changes are sufficiently robust to allow for practical assessment. Practical applications would probably require a single voice measure that is reliable across subjects and situations or, alternately, a battery of voice measures that could be applied to each individual subject and produce a reliable profile of that individual's response to stress.

The present paper reports on a research program that is examining issues related to practical voice assessment. The first part of the program was to identify those candidate voice measures from the available research literature that displayed the greatest promise of responding to psychological stress changes. Eight such measures were identified. The second part of the program was to execute an original laboratory experiment that involved clear physiological changes on the part of the subjects within the type of stress range that might be encountered in routine aerospace activity (as opposed to the higher stress range typically encountered in emergency situations from which much of the scientific voice information has been demonstrated). The experiment employed an aviation-like tracking task, varying both task difficulty and monetary incentives.

The third part of the research program was to automate the eight candidate voice measures and compare their responses within the laboratory data to those of traditional physiological measures such as heart rate. This part of the research program is partially complete, with five of the candidate measures automated, and this paper reports the initial results of this effort.

CANDIDATE VOICE MEASURES

Eight candidate voice measures were determined that, it was believed, showed the greatest promise of responding to psychological stress. The choice of these measures was assisted by a comprehensive literature review completed recently for the Naval Air Test Center (ref. 1) and by the authors' familiarity with recent developments in the voice stress area.

The eight candidate measures are

1) Fundamental frequency (pitch). Under stress, there may be an increase in the fundamental frequency of the voice. Fundamental frequency, which may reflect the physical tension of the vocal muscles, is among the most frequently cited voice indices of stress. In emergency situations an increase in fundamental frequency may be universal (refs. 3, 4 and 5).

2) Amplitude (loudness). Under stress, there may be an increase in the amplitude of the voice. This change would probably reflect an increased air flow through the lungs that often occurs under stress.

3) <u>Speech rate</u>. Under stress, there may be an increase in speech rate. This change would be related to a general speeding up of cognitive and motor processes that often appears under stress.

4) <u>Frequency jitter</u>. Under stress, there may be a decrease in jitter of the voice fundamental frequency. Jitter is the minute variability which occurs in the spacing of the fundamental frequency periods (when measured on a cycle-by-cycle basis). It represents a subtle aspect of audible speech that can be difficult to measure precisely (ref. 6). Lieberman (ref. 7) proposed that jitter decreases in response to psychological stress, and there is recent supporting evidence (ref. 3).

5) <u>Amplitude shimmer</u>. Under stress, there may to be a decrease in shimmer of voice amplitude. Shimmer is the cycle-by-cycle variability in the amplitude pattern (and is the equivalent measure to amplitude that jitter is to frequency). Although no literature relates shimmer to psychological stress, it seems reasonable from theoretical considerations that it might follow a pattern similar to that of jitter.

6) <u>PSE scores</u>. Under stress, there may be an increase in scores determined from the Psychological Stress Evaluator (PSE). The PSE is the best-researched of a series of commercial voice devices sold for lie detection. There is substantial evidence that the PSE is not valid for lie detection (refs. 8 and 9), a questionable application for any stress measure that requires subjective determinations by the person administering the test to infer the presence of lying (ref. 10). However, there is also evidence that the PSE-derived scores may respond to simple manipulations of stress (refs. 11 and 12).

7) <u>Energy distribution</u>. Under stress, there may be an increase in the proportion of speech energy between 500 and 1000 Hz. Scherer (refs. 13 and 14) provides evidence for this effect.

8) <u>Derived measure</u>. Under stress, there may be a reliable increase in a derived measure that statistically combines other measures described above. This approach has been advanced by Brenner (ref. 15), who uses the "improper linear model" of Dawes (ref. 16) to provide a simple statistical combination of component speech measures. In theory, the derived measure should then reflect any unusual changes within the same speaker's voice on one or many component measures. In a recent judicial decision, in the legal case of <u>Hoppie/Gillie v. Cessna</u>, such an approach to voice stress analysis was judged to provide admissible evidence (refs. 17 and 18).

LABORATORY EXPERIMENT

An experiment was designed that, it was hoped, would provide clear physiological differences within the subjects tested. The experiment employed the tracking task of Jex, McDonnell & Phatak (ref. 19), a highly motivating task requiring good reaction time that has been employed extensively in aerospace research. This task can be varied over a wide range of difficulties, and previous literature has suggested physiological changes in response to task loading on measures drawn from heart, respiration, and EMG data (refs. 20 and 21). For the present experiment, monetary incentives were used along with task loading to help guarantee a clear physiological response.

Heart data were obtained from the subjects during the experiment, and excellent voice recordings were obtained of the spoken responses in digital format. Preliminary results available at this time indicate a clear direction for the voice measures that have been tested.

Subjects

Seventeen males, ranging in age from 21 to 35 years old, served as subjects. They were paid $50 plus any monetary incentives won during the experiment.

Procedure

The experiment employed the tracking task of Jex, McDonnell & Phatak (ref. 19) implemented on the Commodore 64 computer. In this task the subject is seated at a CRT display with a manual joystick and attempts to keep a computer-generated triangle at the center of the screen. The triangle moves left and right horizontally in an unpredictable pattern until it touches a left or right boundary on the screen and the trial ends (giving the subject a task similar to balancing a broomstick on a fingertip). A numerical value, the Lambda score, quantifies the mathematical unpredictability of the triangle's gyrations.

Each subject participated at two sessions. At Session 1, the subject was seated in a practice room and trained on the tracking task (25 trials, 10 minute break, 25 trials, ten minute break). At this time subjects performed the "critical" form of the task, in which Lambda was shown on the screen and increased progressively during the trial. The subject attempted to achieve as high a Lambda score as possible before the triangle went out of bounds. To provide speech data, subjects counted aloud on half of the trials. Every ten seconds during the trial, following a computer-generated cueing tone, subjects counted aloud from 90 to 100 as quickly as possible. The counting task was chosen because it causes minimal interference with the tracking task, and the numbers 90 to 100 were chosen because they provide an excellent acoustic pattern with almost continuous voicing.

Following this training, the subject was seated in the laboratory and attached to data recording equipment. Heart rate data were recorded on a multi-channel FM recorder via silver/silver chloride electrode monitors attached to the right and left upper rib areas and base of the neck (the ground electrode). Speech data were recorded via a 1" condensor microphone contained in a custom-modified rubber anaesthesia mask worn by the subject. Speech data were captured digitally in real-time on a laboratory computer at a sampling rate of 10 kHz (the rubber mask also contained a pneumotachograph to measure respiration, and data from this measure are to be described in future papers).

Following a warmup period (ten trials of the "critical" task), subjects performed the "sub-critical" form of the tracking task. In this form the Lambda score, not shown on the screen, was fixed at a specific level of difficulty. The subject's task was to keep the triangle centered for as long as possible up to ninety seconds. On some trials the Lambda score was "easy" (Lambda = 0.9), on some trials "difficult" (Lambda = 90% of the subject's best practice score, median of five trials), and on some trials "moderate" (Lambda = 75% of the subject's best practice score). Each subject performed two trials at each difficulty level. Finally, the subject rested for fifteen minutes, provided baseline measures, and was dismissed. The purpose of Session 1 was training and familiarization, and none of the data collected at Session 1 were analyzed.

367

At Session 2, several days later, the subject was again
seated in the laboratory and attached to data recording
equipment. The subject performed a warmup procedure (ten trials
of the "critical" task). The subject then performed several
trials of the "subcritical" task, both "easy" and "difficult",
and these trials represent the principal source of data for the
experiment. For these trials, the subject was offered monetary
bonuses. On easy trials (Lambda = 0.9) the subject was offered
two dollars if he could complete a successful ninety second trial
within two attempts. All subjects performed perfectly on the
first attempt. On difficult trials (Lambda = 90% of best
practice score) the subject was offered fifty dollars if he could
complete a successful ninety-second trial within two attempts.
Those subjects who failed at this bonus were offered forty-five
dollars and two attempts to complete a slightly less difficult
task (Lambda = 85% of best score). All subjects succeeded by the
end of this second bonus (median Lambda value = 4.2). The order
of easy and difficult presentations was counterbalanced across
subjects.

To complete Session 2, the subject rested for fifteen
minutes and provided baseline measures. The subject was
debriefed, paid, and dismissed.

Data Reduction

An automated program was prepared for data reduction related
to five of the automated speech measures. The extraction of
these parameters was based on algorithms and software developed
by E. Thomas Doherty, Ph.D., of the Speech Research Laboratory,
Veterans Administration Research Laboratory, San Francisco,
California. Dr. Doherty also served as a consultant on this
project, and technical details of the analysis program will be
provided in other reports.

The automated program inputs recorded speech at slow speed,
segmenting it into speech periods and removing the silent periods
between syllables and words. The program outputs automated
measures for five of the candidate speech measures: fundamental
frequency, amplitude, speech rate (ie. total time to speak the
ten numbers), jitter, and shimmer.

Results

Data analysis applied to three trials from Session 2 for each subject: the successful "difficult" trial on which the subject won $50 or $45; the successful "easy" trial on which the subject won $2; and a baseline trial on which the subject simply counted. Speech data on each trial consisted of nine repetitions of the numbers 90 to 100.

Figure 1 displays heart rate data. Average heart rate was 83 bpm on the baseline trial, 88 bpm on the easy trial, and 100 bpm on the difficult trial ($F_{(2/32)} = 22.1$, $p < .001$). An analysis-of-variance test proved highly significant for the overall difference between difficult and easy ($F_{(1/32)} = 21.2$, $p < .001$), and 16 of the 17 subjects showed a higher average heart rate on the difficult treatment than on the easy treatment (sign test: $p < .001$). Based on the heart rate data, then, the experiment produced a clear physiological response against which the voice measures can be compared.

Speech data are summarized in Tables 1 and 2 and in Figures 2, 3, and 4. The analysis-of-variance values reported in Tables 1 and 2 are for differences between the treatment means (a more complete analysis, treatment x time, has not been completed). The second column of Table 2 ("Number of subjects with predicted effect") represents a sign test.

Amplitude displayed a highly significant relation to the task and, as shown in Figure 2, provided a pattern resembling that of heart rate. Average amplitude increased between the easy and difficult treatments by a magnitude of about 0.07 volts, a change that was clearly measurable but that would be virtually impossible to recognize in normal conversation. Fundamental frequency also increased in response to the task, providing a pattern of results less robust than that of amplitude. Average fundamental frequency varied between the easy and difficult treatments by a magnitude of about 2 Hz., a change that is also negligible in normal conversation.

The speech rate measure provided a marginally significant discrimination of the three treatments. Speech rate also showed the highest consistency across subjects of any of the speech measures.

The jitter measure responded in the predicted direction, but to a marginal degree that produced little statistical effect. This measure is of theoretical interest but, pending the results of a complete analysis, does not appear to respond to the type of stress present on this task. Shimmer also responded with marginal effect, but showed a consistency across subjects that suggests a need for further study.

CONCLUSIONS

Previous literature has reported increases in fundamental frequency, amplitude, and speech rate in the voices of speakers involved in extreme levels of stress (refs. 3, 4, and 5)(and these changes are among the major components of screaming). What seems remarkable about the present results is that the same changes appear to occur in a regular fashion within a more subtle level of stress that may be characteristic, for example, of routine flying situations. This evidence adds confidence that these changes reflect some valid underlying physiological response of the human speech system.

The results of our experiment replicate exactly those reported recently by Griffin & Williams (ref. 22). Working in an aircraft simulator setting, they found that increases in speech amplitude, fundamental frequency, and speech rate appeared in the subjects' speech in response to increased workload demands. The combined evidence of the experiments helps establish these three voice measures as parameters for aerospace applications.

In our research, none of the individual speech measures performed as robustly as did heart rate. An area of active future interest is to develop a single derived speech measure, drawing information from several component speech aspects, and to compare the performance of this measure with that of a measure such as heart rate. Another area of future interest is the possibility of developing a convenient and even real-time assessment technique, especially given the current explosion in automated speech processing technology. Voice stress analysis is maturing as a research area, and we urge our colleagues to consider voice response in their thinking about mental-state estimation.

REFERENCES

1. Biers, D. W.: Voice Stress as a Workload Measure: Review of the Literature and Recommendation for Further Study. Final Report, Contract N00421-83-D-0027, Naval Air Test Center, Patuxent River, Maryland, March 1984.

2. Cannings, R.: Speech Patterns and Aircraft Workload. In: Survey of Methods to Assess Workload. NATO AGARDOGRAPH 246, 115-129, August 1979.

3. Brenner, M., Shipp, T., Doherty, E. T., & Morrissey, P.: Voice Measures of Psychological Stress--Laboratory and Field Data. In: I. R. Titze & R. C. Scherer (Eds.) Vocal Fold Physiology, Biomechanics, Acoustics, and Phonatory Control. Denver: The Denver Center for the Performing Arts, 1985.

4. Kuroda, I., Fujiwara, O., Okamura, N., & Utsuki, N.: Method for Determining Pilot Stress Through Analysis of Voice Communication. Aviation, Space, and Environmental Medicine, 47, 1976, 528-533.

5. Williams, C. E., & Stevens, K. N.: Vocal Correlates of Emotional States. In J. K. Darby (Ed.). Speech Evaluation in Psychiatry. New York: Grune and Stratton, 1981.

6. Doherty, E. T., & Shipp, T. Tape Recorder Effects on Jitter and Shimmer Extraction. Unpublished Manuscript, Speech Research Laboratory, Veterans Administration Medical Center, San Francisco, California, 1987.

7. Lieberman, P.: Perturbation in Vocal Pitch. Journal of the Acoustical Society of America, 33, 1961, 597-603.

8. Brenner, M., & Branscomb, H. H.: The Psychological Stress Evaluator; Technical Limitations Affecting Lie Detection. Polygraph, 8, 1979, 127-132.

9. Horvath, F.: Detecting Deception: The Promise and the Reality of Voice Stress Analysis. Journal of Forensic Sciences, 27, 1982, 340-351.

10. Lykken, D. T.: Psychology and the Lie Detector Industry. American Psychologist, 29, 1974, 725-739.

11. Brenner, M., Branscomb, H. H., & Schwartz, G. E.: Psychological Stress Evaluator--Two Tests of a Vocal Measure. Psychophysiology, 16, 1979, 351-357.

12. Schiflett, S. B., & Loikith, G. J.: Voice Stress Analysis as a Measure of Operator Workload. Patuxent River, Maryland: Naval Air Test Center, TM-79-3-SY, March 1980.

13. Scherer, K. R.: Speech and Emotional States. In J. K. Darby (Ed.). Speech Evaluation in Psychiatry. New York: Grune and Stratton, 1981.

14. Scherer, K. R.: Vocal Indicators of Stress. In J. K. Darby (Ed.). Speech Evaluation in Psychiatry. New York: Grune and Stratton, 1981.

15. Brenner, M.: Analyzing Voice Tapes for Psychological Stress. Paper presented at the Annual Southern Methodist University Air Law Symposium, Dallas, Texas, 1986.

16. Dawes, R. M.: The Robust Beauty of Improper Linear Models in Decision Making. American Psychologist, 34, 1979, 571-582.

17. Brenner, M.: Affidavit in Hoppie/Gillie v. Cessna. United States District Court, Missoula, Montana, May 1986.

18. Hollien, H.: Affidavit in Hoppie/Gillie v. Cessna. United States District Court, Missoula, Montana, June 1986.

19. Jex, H. R., McDonnell, J. D., & Phatak, A. V.: A "Critical" Tracking Task for Manual Control Research. IEEE Transactions on Human Factors in Electronics, HFE-7, 1966, 138-145.

20. Derrick, W. L., & Wickens, C. D.: A Multiple Processing Resource Explanation of the Subjective Dimentions of Operator Workload. Technical Report EPL-84-2/ONR-84-1, Engineering-Psychology Research Laboratory, University of Illinois at Urbana-Champaign, February 1984.

21. Jex, H. R., & Allan, R. W.: Research on a New Human Dynamic Response Test Battery. Paper presented at the Sixth Annual Conference on Manual Control, Wright-Patterson AFB, Ohio, April 1970.

22. Griffin, G. R., & Williams, C. E.: The Effects of Different Levels of Task Complexity on Three Vocal Measures. Aviation, Space, and Environmental Medicine, In Press (December, 1987).

Table 1. Differences between the treatment means for the five voice measures (analysis of variance).

|	F (2/32)
Fundamental Frequency	7.1**
Amplitude	10.2***
Speech Rate	3.1*
Jitter	0.1
Shimmer	1.3

* p<.10
** p<.01
*** p<.001

Table 2.   Differences between the easy and difficult treatment means for the five voice measures (analysis of variance/sign test).

	F (1/32)	Number of subjects with predicted effect
Fundamental Frequency	2.9*	10/17
Amplitude	5.0**	13/17**
Speech Rate	2.5	14/17***
Jitter	0.1	9/17
Shimmer	0.7	13/17**

* p<.10
** p<.05
*** p<.005

Figure 1

Difficult
Easy
Baseline
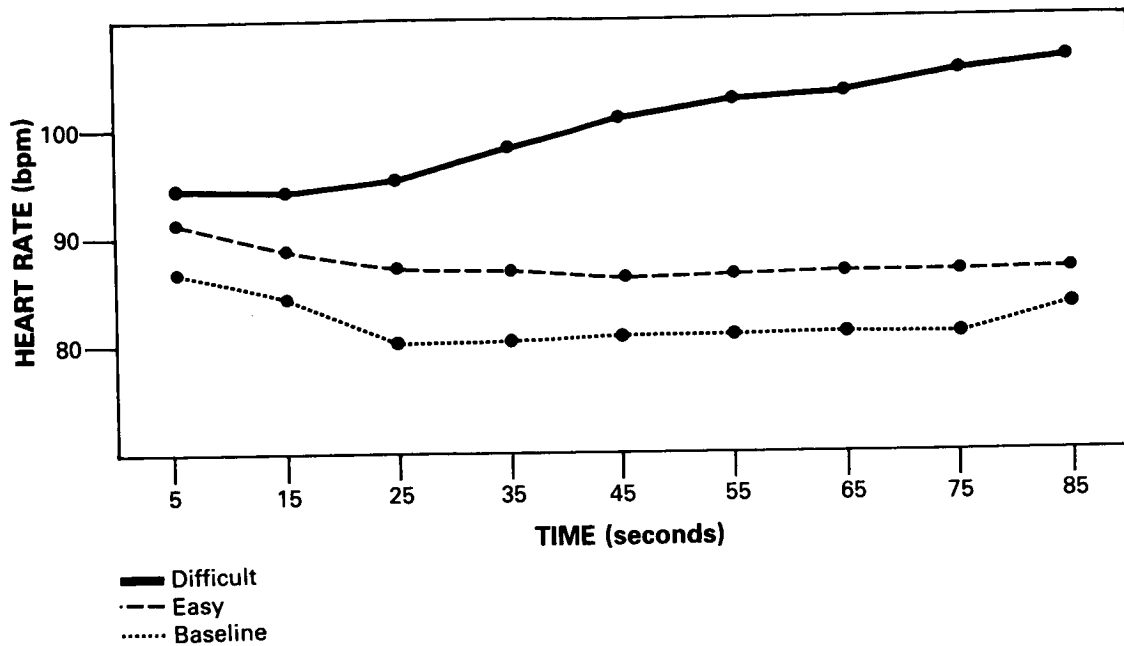
1. Average heart rate over a ninety-second trial as a function of the experimental manipulation.

Figure 2

Difficult
Easy
Baseline
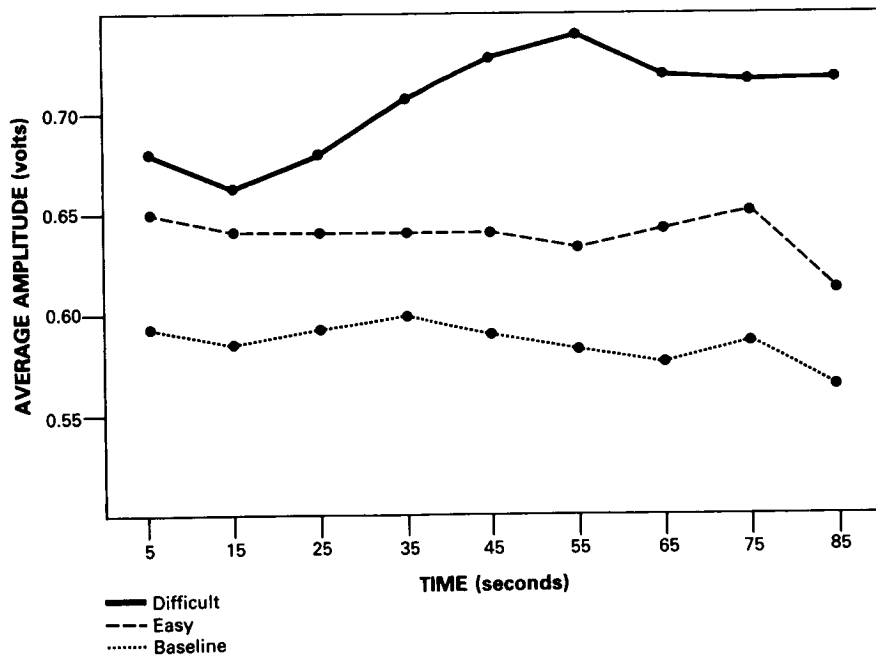
2. Average speech amplitude over a ninety-second trial as a function of the experimental manipulation.

Figure 3

Difficult
Easy
Baseline
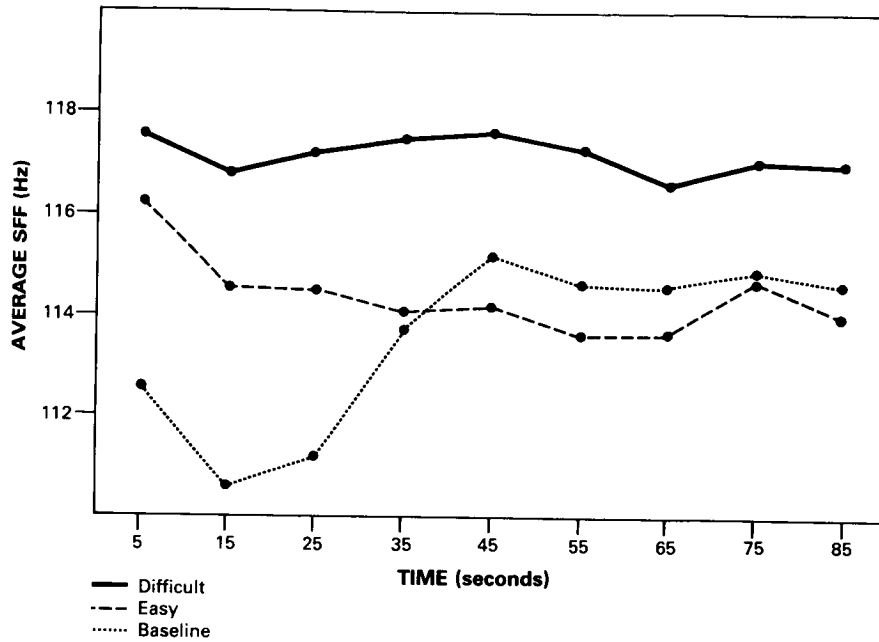
3. Average speech fundamental frequency over a ninety-second trial as a function of the experimental manipulation.

Figure 4

Difficult
Easy
Baseline
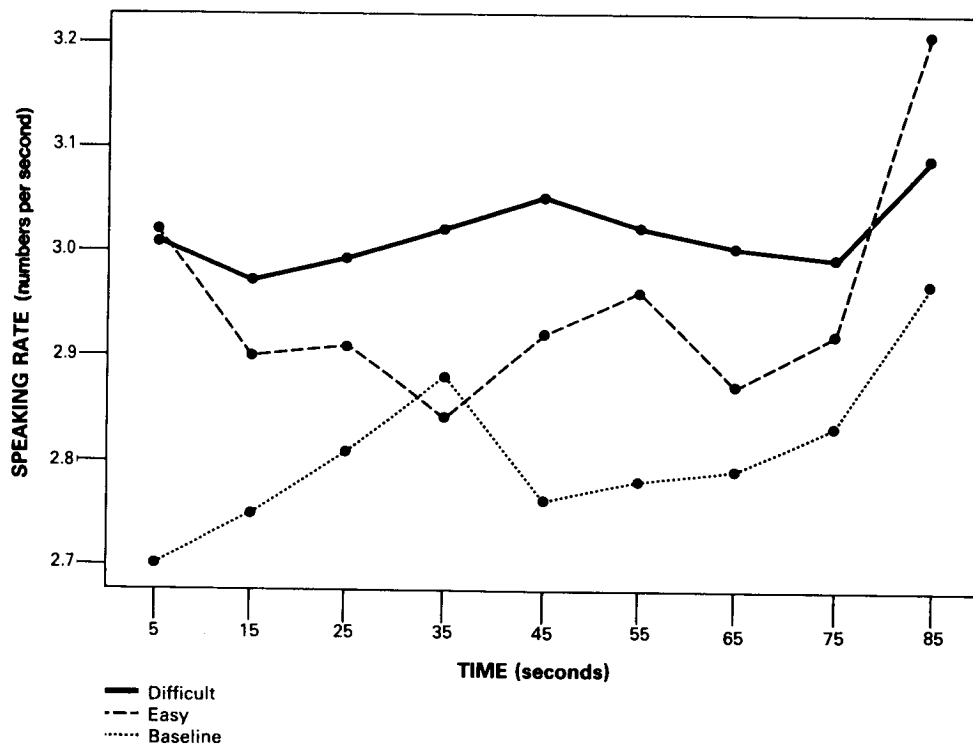
4. Average speech rate (while speaking) over a ninety-second trial as a function of the experimental manipulation. The subject recited the numbers 90-100 every ten seconds.

376