Audio Engineering Society

Convention Paper

Presented at the 134th Convention
2013 May 4–7 Rome, Italy

Pilot Workload and Speech Analysis A Preliminary Investigation

Rachel M. Bittner[1], Durand R. Begault[2], and Bonny R. Christopher[3]

[1] *Courant Institute of Mathematical Sciences, New York University, New York City, NY, 10012, USA*

[2] *Human Systems Integration Division, NASA Ames Research Center, Moffett Field, CA, 94035, USA*

[3] *San José State University Research Foundation, NASA Ames Research Center, Moffett Field, CA, 94035, USA*

Correspondence should be addressed to Durand Begault (durand.r.begault@nasa.gov)

ABSTRACT

Prior research has questioned the effectiveness of speech analysis to measure the stress, workload, truthfulness, or emotional state of a talker. The question remains regarding the utility of speech analysis for restricted vocabularies such as those used in aviation communications. A part-task experiment was conducted in which participants performed Air Traffic Control read-backs in different workload environments. Participant's subjective workload and the speech qualities of fundamental frequency (F_0) and articulation rate were evaluated. A significant increase in subjective workload rating was found for high workload segments. F_0 was found to be significantly higher during high workload while articulation rates were found to be significantly slower. No correlation was found to exist between subjective workload and F_0 or articulation rate.

1. INTRODUCTION

Research has questioned the effectiveness of the use of speech analysis to measure stress, workload, truthfulness or other factors related to the emotional state of a talker for well over half a century. The use of voice stress analysis as a form of lie detection in commercial products has been disparaged and for the most part is not taken seriously in the scientific community [1]. This includes measures of low-frequency "micro-tremors" in the laryngeal muscles as indicators of psychological stress related to deception, analogous to polygraph measures [2].

The relationship between speech measures and psychological stress caused by high-stress activities has received more scientific attention, although these studies have not defined "workload" or "stress" in a consistent manner. Cain's (2007) review asserts that no commonly accepted, formal definition of workload exists, but states, "workload can be characterized as a mental construct that reflects the mental strain resulting from performing a task under specific conditions" [3]. Brenner et. al (1994) defines workload to mean changes in task demands and incentives, and stress as psychological changes that result from workload demands, which may be observed from changes in physiological measures [4]. Ruiz et al. (1990) defines workload as pathological elements characterized by physical fatigue and psychological stress that are carried in the speech signal [5]. Literature reviews indicate that physiological measures that attempt to unambiguously identify a "tell tale" sign of workload, such as heart rate or evoked potential monitoring, have generated mixed results overall, and are not generalizable to all persons [6].

Investigations of the relationship between workload and its effects on speech in an aeronautic context are exemplified by Brenner and Shipp (1987), who measured stress in terms of heart rate and found significant correlation with both speech rate and fundamental frequency [7]. However, other studies have shown measures such as fundamental frequency are inconclusive with regards to stress or workload. Hecker et al. (1968) examined task-induced stress and found that "the acoustical effect of stress for one individual may be quite different from those for another individual," concluding that the manifestations of stress were well defined only for some individuals [8]. Ruiz et al.'s (1990) review of the literature concluded that a single parameter analysis, for example fundamental frequency, was insufficient, but that multi-dimensional measures would be promising [5].

The question remains about the utility of speech analysis for restricted vocabularies such as those used in aviation communications. An exploratory study by one of the authors examined the differences in fundamental frequency and articulation rate among single pilots flying very light jets in high and low workload contexts, based on both read back and other communications [9]. No significant differences were found, likely due to the minimal number of subjective workload ratings obtained.

The goal of this study was to determine whether speech analyses of fundamental frequency and articulation rate of pilot "read back" of air traffic control (ATC) commands could be used as a basis for detecting changes in the pilot's workload. "Read back" is a characteristic aviation voice communication technique between pilots and air traffic controllers, where the essential details of a command (e.g., flight level or direction of turn) are repeated by the receiving party. For the present study, workload was defined as the level of mental and physical work needed to maintain performance in a communication read back task, as confirmed by the participant's subjective workload rating. We aimed to achieve the following objectives:

1. Ensure a difference in subjective workload (participant's perceived level of workload) and task workload (speed of distractor messages presented) exists.
2. To investigate the influence of task workload on articulation rate.
3. To investigate the influence of task workload on F_0.
4. To examine if a linear relationship exists between subjective workload and articulation rate.
5. To examine if a linear relationship exists between subjective workload and F_0.

We predicted that subjective workload ratings would increase with an increase in task workload based on an experimental manipulation of a distractor task rate. Likewise we predicted that the average F_0 and articulation rate for a read back phrase would be significantly different as a function of task workload. No direction was specified for the change in F_0 and articulation rate. We also predicted that if a significant difference existed in subjective workload as a function of the experimental manipulation, a linear relationship would also exist between subjective workload and F_0 and articulation rate.

Fig. 1: Virtual (iPad) button display for answering incoming ATC messages. (The second button is yellow to indicate it is active)

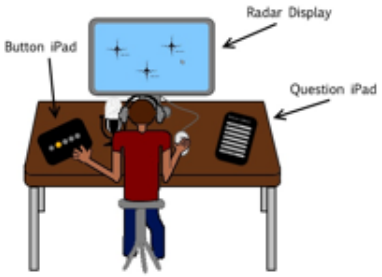

Fig. 2: Experimental Setup

2. METHOD

2.1. Participants

Seventeen participants were recruited for the study (9 men and 5 women, age range 18 to 33) via the Subject Recruitment Office of the San Jose State University Foundation. The experiment was approximately 60 minutes including a training and familiarization session. Participants were paid a nominal amount for participation in the study. The study protocol was covered under an omnibus Human Research Institutional Review Board approval from NASA Ames Research Center.

2.2. Experimental Setup and Equipment

The experiment was performed in a sound proof booth where participants heard all audio on open circumaural dynamic headphones (Sennheiser HD 595). Ambient aircraft noise was played through the headphones throughout the course of the experiment to simulate pilot aural conditions. A high quality speech synthesizer was used to deliver incoming messages. Two portable touch screen interfaces (Apple iPad II) were used; one to answer ATC calls using a virtual button display (see Figure 1) and another in which participants were prompted to provide subjective ratings on a seven-point Likert scale. A computer display was used to present pseudo-radar information to the participant and a computer mouse was used to select direction in the distractor task. All communication was recorded using a high quality microphone (AKG C414 B-ULS). An illustration of the experimental setup is shown in Figure 2.

2.3. Procedure

Prior to the start of the experiment, participants were trained to use a restricted aviation-specific vocabulary. Participants were asked to play the role of "pilot" and were tasked with answering incoming messages from ATC, confirming instructions by radioing ATC, and indicating nearby aircraft positions on flight radar.

Participants were presented incoming messages via the left side iPad with five interactive buttons. Each button represented an open line in which the ATC might call. Active lines with incoming messages were presented in yellow and non-active lines were shown in grey. When a line appeared yellow the participant selected the line by pressing on the touch screen display, this allowed the ATC message to be played for the participant.

Two types of ATC messages were given to the pilot, priority and distractor. Priority messages were addressed specifically to the participant's aircraft, and participants were asked to respond to priority messages by providing ATC a read-back of the instruction. For example, a priority ATC message could be "United 972, Saint Louis center. Climb to flight level 380." and the participants response would be, "Saint Louis center this is United 972. Climbing to flight level 380."

Distractor messages were instructions given to other aircraft presented on the radar. These messages did not require a verbal response from the participant; however, participants were asked to acknowledge the message by selecting the appropriate arrow on the flight response indicator on the computer screen using a mouse. The instructions needed to identify which arrow to select in the distractor task were delivered at the end of each message, forcing the participants to listen to the entire message. For example, in response to the message "Jet Blue 877, San Francisco Center. Turn right heading 90." the

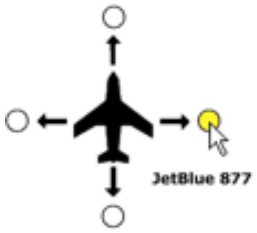

Fig. 3: Example user response to the distractor message "*Jet Blue 877, San Francisco Center. Turn right heading 90.*"

participant would click the right arrow as shown in Figure 3.

Participants were asked to provide a subjective workload rating once per minute. Participants were prompted via the right side touch screen to "Rate your workload". A seven point interactive Likert scale was presented with one representing "very low" workload and seven being "very high" workload.

The experiment was broken into two 20 minute, 30 second blocks. Within each block were two 10 minute segments separated by 30 seconds. For all segments, participants received one priority message per minute. During high workload segments, participants received 10 distractor messages per minute, and during low workload segments, 5 distractor messages per minute. A 10 minute break was provided between blocks. Tasks in both segments and blocks were identical however the order of workload was varied.

2.3.1. Design

The independent variable was task workload, at two levels: high (10 distractor messages per minute) and low (five distractor messages per minute). The dependent variables were subjective rating, F_0, and articulation rate. Participant's F_0 and articulation rate were derived using PRAAT software and custom algorithms [10, 11] for each priority readback [3]. Average fundamental frequency was measured in Hz and articulation rate was calculated by the number of syllables divided by phonation time (i.e., the total speaking time from the start to the end of read back utterance, not including pauses in speaking). An example of the measurements taken for one subject is shown in Figure 4.

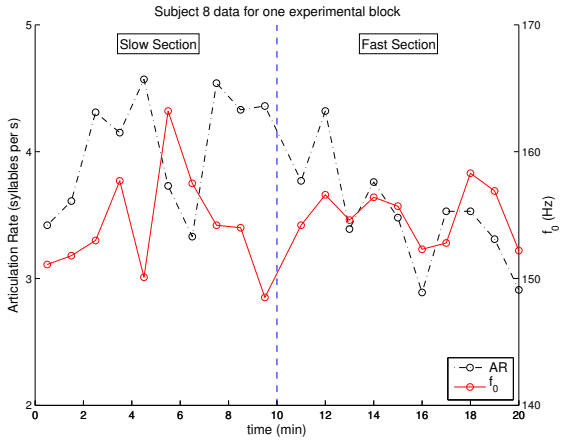

Fig. 4: Subject 8's raw data for one block.

3. RESULTS

Prior to statistical analyses, data from three participants were removed because they did not complete the task correctly. Outliers beyond two standard deviations were removed prior to averaging F_0 and articulation rate.

In order to ensure that participants were sensitive to the low and high levels of task workload manipulation, a Wilcoxon signed rank test was conducted on their subjective workload ratings. Participants reported a significant increase in perceived workload during high workload segments ($Mdn = 4.54$) as compared to low workload segments ($Mdn = 2.00$), $z = 4.3734$, $p < .0005$, $r = 0.580$.

To test our prediction that F_0 would be different under high and low workload, a paired samples t-test was used to investigate whether there was a statistically significant mean difference of participant's average F_0 between the two workload segments. Two outliers were detected and removed from the analysis. The data was normally distributed, as assessed by the Shapiro-Wilk test ($p = .216$). Participant's average F_0 was found to be significantly higher during high workload segments ($M = 159.853$, $SD = 48.723$) than low workload segments ($M = 158.319$, $SD = 47.753$), a statistically significant mean increase of 1.534 Hz, 95% CI [0.61872 to 2.45025], $t(11) = 3.688$, $p < .01$, $d = 1.065$.

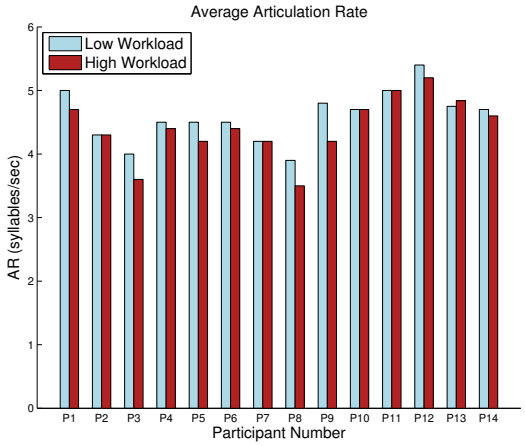

Fig. 5: Average AR of each subject in each condition.

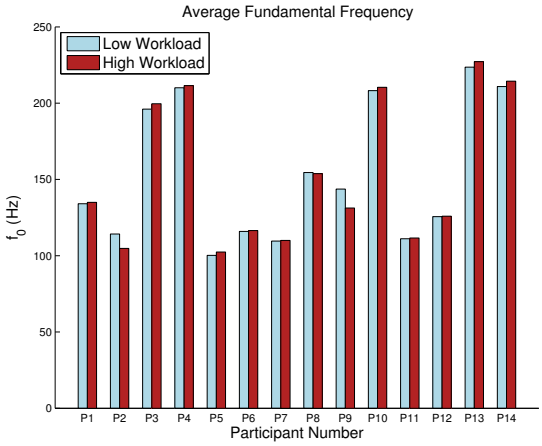

Fig. 6: Average F_0 of each subject in each condition.

To investigate our prediction that a linear relationship would exist between our speech variables and subjective rating (based on the results of the previous tests) we planned to conduct two Pearson product-moment correlations. In order to meet the criteria of the statistical test (i.e., continuous variables), delta scores (effect size) were used in lieu of raw data. However, preliminary analyses of data plots did not indicate a linear or monotonic relationship to exist between the effect size of subjective rating and F_0 or the effect size of subjective rating and articulation rate.

4. DISCUSSION

With this experiment we were able to show that participants perceived a difference in our high and low workload sections, and these differences were demonstrated in the speech variables measured. Participant's perceived workload increased as a result of increasing the number of distractor messages indicating that there was an actual difference in workload between the trials. As predicted, we found a significant 1.5% decrease in participant's F_0 when participants took part in the slow trial as compared to the fast trial. Also confirming our predictions was a 0.14% increase in articulation rate during the slow trials. We did not find a relationship between the speech measures researched and the participant's subjective ratings of workload. This was likely due to the variability of the subjective responses across the task.

We observed that the use of a more controlled and restricted aviation vocabulary restricted the range of fundamental frequency and articulation rate differences compared to those that occur in normal everyday speech, in an idiosyncratic manner for each talker. In a study using a counting task, Brenner et al. [4] found a ∼2 Hz increase in fundamental frequency, which compares in magnitude to the 1.5 Hz increase found here. They observed that, while "the degree of change would be difficult to recognize in normal conversation" the measured effect deserved "special attention for practical aerospace applications." While we found that fundamental frequency increased with workload, similar to [4], that study also showed an increase in articulation rate with increased workload for most subjects (about a 0.5 syllable increase/sec), whereas the current study found a decrease in articulation rate. The decrease in articulation rate with increased workload found here is consistent with several studies focused on the effects of cognitive load, where the experimental manipulation involved increasingly difficult speaking tasks [12].

Our results also suggest that, notwithstanding the controlled aviation vocabulary, each person's voice responds differently to stress in the given context.

For example, three of the subject's fundamental frequencies decreased slightly under stress, and two additional subjects had a slight increase in articulation rate, contrary to the direction of the mean. These individual differences can still be used as an indicator of stress if the direction of change is normalized. Future research might investigate the advantages of a more complex analysis model, tailoring speech stress analysis to each individual's characteristic response for a given task.

5. REFERENCES

[1] Eriksson, A. and Lacerda, F. (2007). Charlatanry in forensic speech science: A problem to be taken seriously. Int. Journal of Speech, Language and the Law. vol. 14, pp. 169-193.

[2] Hovarth, F. (1982). Detecting Deception: The promise and the Reality of Voice Stress analysis. Journal of Forensic Sciences, vol. 27, pp. 340-351.

[3] Cain, B. (2007). A review of the mental workload literature. Defence Research and Development Canada. Report no. RTO-TR-HFM-121.

[4] Brenner, M., Doherty, E. T., and Shipp. T. (1994). Speech measures indicating workload demand. Aviation, Space and Environmental Medicine pp. 21-26.

[5] Ruiz, R., Legros, C., and Guell, A. (1990). Voice analysis to predict the psychological or physical state of a speaker. Aviation, Space and Environmental Medicine, vol. 61. pp. 266-271.

[6] Casner, S. M., and Gore, B. F. (2010). Measuring and Evaluating Workload. A Primer. NASA Technical Memorandum no. 2010-216395.

[7] Brenner, M. and Shipp, T. (1987). Voice stress analysis. In Constuck, J. R. (ed.) Mental State Estimation. NASA Conference Publication 2504.

[8] Hecker, M. H. L., Stevens, K. N., von Bismarck, G.D. and Williams, C. E. (1968). Manifestations of task-induced Stress in the Acoustic Speech Signal. Journal of the Acoustical Society of America, vol. 44, pp. 993-1001

[9] Burian, B. K., Pruchnicki, S., Rogers, J., Christopher, B., Williams, K., Silverman, E., Drechsler, G., Mead, A., Hackworth, C., and Runnels, B. (2012). Single pilot workload management in entry level jets. Final Report. http://humansystems.arc.nasa.gov/flightcognition/publications

[10] Boersma, P. and Weenink, D. (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3.41. http://www.praat.org/

[11] De Jong, N. H. and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. Behavior research methods, vol. 41, pp. 385 - 390.

[12] Berthold, A., and Jameson, A. (1999). Interpreting symptoms of cognitive load in speech input. In J. Kay (Ed.), User Modeling: Proceedings of the Seventh International Conference, UM99. Vienna, New York: Springer Wien New York, 1999.