USAARL-TECH-FR--2021-22

Non-Invasive Physiological Biomarkers of Cognitive Fatigue in a Virtual Reality Simulated, Rotary-Wing Flight Environment

Gregory Ciccarelli, Hrishikesh M. Rao, Christopher J. Smalt,
Harvey Edwards, Daryush Mehta, Hayley Reynolds, Kara Cave,
& Thomas Quatieri

Notice

Qualified Requesters

Qualified requesters may obtain copies from the Defense Technical Information Center (DTIC), Fort Belvoir, Virginia 22060. Orders will be expedited if placed through the librarian or other person designated to request documents from DTIC.

Change of Address

Organizations receiving reports from the U.S. Army Aeromedical Research Laboratory on automatic mailing lists should confirm correct address when corresponding about laboratory reports.

Disposition

Destroy this document when it is no longer needed. Do not return it to the originator.

Disclaimer

The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation. Citation of trade names in this report does not constitute an official Department of the Army endorsement or approval of the use of such commercial items.

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

This material is based upon work supported by the Department of the Army under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of the Army.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
07-07-2021	Final	1 June 2018 - 1 June 2021

4. TITLE AND SUBTITLE

Non-Invasive Physiological Biomarkers of Cognitive Fatigue in a Virtual Reality Simulated, Rotary-wing Flight Environment

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

MOMRP 20470

6. AUTHOR(S)

Ciccarelli, G.[1], Rao, H.[1], Smalt, C.[1], Edwards, H.[1], Mehta, D.[1], Reynolds, H.[1], Cave , K.[2], & Quatieri, T.[1]

5d. PROJECT NUMBER

240

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

U.S. Army Aeromedical Research Laboratory
P.O. Box 620577
Fort Rucker, AL 36362

8. PERFORMING ORGANIZATION REPORT NUMBER

USAARL-TECH-FR--2021-22

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

MIT Lincoln Laboratory
5 Forbes Rd
Lexington, MA 02421

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

[1]Human Health and Performance Systems Group, MIT Lincoln Laboratory; [2]U.S. Army Aeromedical Research Laboratory

14. ABSTRACT

Timely and accurate monitoring of aviator cognitive workload offers a means to identify and mitigate aviation mishaps. This study examined a proof-of-concept for a non-invasive, multi-modal platform to quantify the relationship between physiological indicators of pilot fatigue and operational performance in simulated flight tasks. Seven participants (two females) varying from no piloting experience to a commercially rated pilots, completed a 90-min repetitive flight traffic pattern. The psychomotor vigilance task (PVT), flight performance data, and thirteen physiological sensing modalities monitored fatigue and performance. Post-flight PVT reaction times were longer as compared to pre-flight baselines. Further, vocal biomarkers analyses support the existence of a coupling between fine motor aspects of speech production and flight performance in more experienced aviators. These preliminary results show the potential for speech to be used to predict real-time flight performance. In future work, addition sensing modalities (e.g., eye tracking, electrocardiogram, electrodermal activity, and torso accelerometry) are to be included into the analysis.

15. SUBJECT TERMS

cognitive workload, vocal biomarker, multi-modal physiological workload, pilot fatigue

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Loraine St. Onge, PhD
UNCLAS	UNCLAS	UNCLAS	SAR	37	19b. TELEPHONE NUMBER *(Include area code)* 334-255-6906

This page is intentionally blank.

Acknowledgements

The authors would like to thank Colonel Ian Curry (United Kingdom [UK] Army) and Adam Lammert (Worcester Polytechnic Institute) for their involvement in the initial project discussions. We also would like to thank Major Murphy (U.S. Army Aeromedical Research Laboratory [USAARL]), Tom Hirsch (East Coast Aero Club), and Major Kyle McAlpin (U.S. Air Force) for discussion involving the protocol design and flight patterns. Finally, we would like to thank William Irvin (Oak Ridge Institute for Science and Education [ORISE], USAARL) for evaluating our Lab Streaming Layer hardware/software setup.

This research was supported in part by an appointment to the Postgraduate Research Program at the U.S. Army Aeromedical Research Laboratory administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Army Medical Research and Development Command.

This page is intentionally blank.

Table of Contents

Table of Contents (Continued)

List of Figures

Table of Contents (Continued)

Page

List of Tables

This page is intentionally blank.

Introduction

The primary objective of this program is to develop and optimize non-invasive biomarkers (e.g., vocal, facial) for quantitative, non-invasive measurements of pilot fatigue in operational environments. The required goals to achieve that objective are:

1. Development and validation of a multimodal assessment platform comprising non-invasive cognitive and psychophysiological tools, as well as strategies for characterization and evaluation of cognitive fatigue and performance; and

2. Identification and implementation of algorithms for extracting behavioral biomarkers and detecting and monitoring fatigue and performance of U.S. Army pilots in operational situations that are extensible to dismounted Warfighters as well.

These goals pave the way toward developing non-invasive sensing modalities that have an identified roadmap toward scaling into a field-deployable fatigue and performance monitoring solution for other Military Operational Medical Research Program (MOMRP) environments.

The ability to concurrently collect flight and human performance metrics would quantify connections between pilot fatigue and operational safety concerns. Specifically, we are interested in testing the following hypothesis with this capability: vocal measures enable direct, quantitative, performance prediction.

Literature Review and Background

Early, accurate detection of diminished or impaired cognitive performance, regardless of etiology, can help reduce the occurrence of accidents and injuries, facilitate timely intervention, and inform treatment/rehabilitation efforts in the recovery period. Assessment and detection of cognitive fatigue is especially important for pilots, where the impact of decreased performance can have dramatic consequences (e.g., the well-known gear-up landing of a C-17 Globemaster in Bagram in January of 2009). In addition, the ability to identify pilots at risk for decreased performance under such conditions can provide an opportunity for intervention, appropriate countermeasures, and enhanced training to reduce impairment and recovery time and to prevent future adverse outcomes. U.S. Army pilots, in particular, are subject to multiple stressors that contribute to cognitive and physiological fatigue within missions that provide neither the operational tempo (OPTEMPO) nor environment conducive for current laboratory-based fatigue and vigilance assessment protocols such as the Psychomotor Vigilance Test (PVT).

The present study involves the development and validation of an unprecedented multimodal platform of non-invasive assessment tools for cognitive performance, associated cognitive status during recovery, and return-to-duty decision making in Army pilots and other military Service Members. Numerous factors, stemming from reversible and irreversible causes, can degrade cognitive readiness and influence recovery in healthy Service Members. These can include fatigue due to physical exertion or sleep loss, and sustained psychological and cognitive stress. All factors, when occurring within operational settings, are exacerbated by sleep and nutritional restrictions, increasing individual risk for accidents, illness, and injuries.

There is mounting evidence that the acoustic speech signal, facial expressions, and physiological measurements are rich with information about an individual's cognitive state, and have been linked with drowsiness, depression, post-traumatic stress disorder, and a variety of neurological and motor control disorders. Therefore, behavioral biomarkers taken from the voice hold promise to be simple, non-invasive indicators and predictors of cognitive fatigue and performance. Speech movements are complex motor activities requiring precise neural timing and coordination. Changes in cognitive fatigue level may predicatively alter this complex motor activity. Information extracted from speech recordings should facilitate continuous monitoring of cognitive performance in both laboratory and operational settings, and provide a replacement for the relatively time- and attention-consuming PVT used in laboratory protocols.

Motivation.

The ease of obtaining vocal features (e.g., voice via helmet microphones, wearable bone conduction and/or skin contact microphones, or via mobile tablets or smartphones) greatly increases global accessibility to an automated method for cognitive assessment. Certain vocal features have been shown to change with a subject's mental and emotional state under numerous conditions including cognitive load and neurological conditions. For voice, these features include characterizations of prosody (e.g., fundamental frequency and speaking rate), spectral representations (e.g., mel-cepstra), and glottal excitation flow patterns, such as flow shape, timing jitter, amplitude shimmer, and aspiration (Ozdas et al., 2004; Darby et al., 1984; Fava & Kendler, 2000).

Approach to Analysis.

Massachusetts Institute of Technology Lincoln Laboratory (MIT LL) seeks an approach that is easy to administer, sensitive, and non-invasive, allowing early detection of effects associated with fatigue, as well as tracking the progression of fatigue and performance over time. The approach proposed here uses MIT LL vocal biomarkers (patents filed) which satisfy these criteria, requiring simply a microphone and digital storage device, in a system having been demonstrated as a sensitive measurement tool for numerous neurological disorders and stresses (MIT LL publications on speech: [Yu et al., 2014; Yu et al., 2015; Williamson et al., 2014; Williamson et al., 2013; Horwitz et al., 2013; Talkar et al., 2020; Quatieri et al., 2017]).

We begin with standard "low-level" features and build upon these to obtain "high-level" timing- and coordination-based features. For voice, the low-level features are phoneme boundaries, formant (vocal tract resonance) tracks, delta mel-cepstra coefficients (spectral dynamics), and creakiness (vocal-fold irregularity). The high-level timing features from phonemes include phoneme-based measures of rate, duration, pitch dynamics, and pause information. The high-level coordination features for all modalities are based on eigenspectra analysis of covariance, correlation, and coherence matrices that are constructed from sets of low-level features. Various subsets of these features have been used at MIT LL in cognitive stress (Quatieri et al., 2017) and neuro-cognitive contexts such as in detection of depression, Parkinson's disease, cognitive impairment, and traumatic brain injury (Krishnan et al., 2012; Malyska et al., 2005; Williamson et al., 2011; Williamson et al., 2013), thus perhaps forming a common feature basis for neurocognitive change.

MIT LL Preliminary Cognitive Fatigue, Load, and Performance work

Prior work at MIT LL has established that differences in cognitive load can be detected from voice and face measurements and compare with electroencephalogram (EEG) signal analysis. Participants engaged in the primary task of verbally recalling sentences with varying levels of cognitive load, as determined by the number of digits being held in working memory (Levitt, 1971; Le et al., 2009; Harnsberger et al., 2008). Specifically, a single trial of the auditory working memory task comprised: the subject hearing a string of digits, then hearing a sentence, then waiting for a tone eliciting spoken recall of the sentence, followed by another tone eliciting recall of the digits. This task was administered with three difficulty levels, involving 108 trials per level. The same set of 108 sentences was used in each difficulty level. The order of trials (sentences and difficulty level) was randomized. The multi-talker PRESTO sentence database was used for sentence stimuli (Park et al., 2010). We recorded 17 subjects but used 11 subjects from whom robust recordings were obtained in all three modalities.

Figures 1 and 2 summarize the results (detection versus false alarm) for each modality alone and in combination. In Figure 2, we see a comparison of receiver operative characteristics (ROCs) across each modality alone. We observe that the EEG-based detector converges to area under the curve (AUC) = 0.99 after 60 trials. The audio and video modalities converge more slowly to AUC = 0.89 and 0.84 individually, and 0.93 in combination. Finally, combining all three modalities converges to near perfect performance, with an AUC of 1.00.

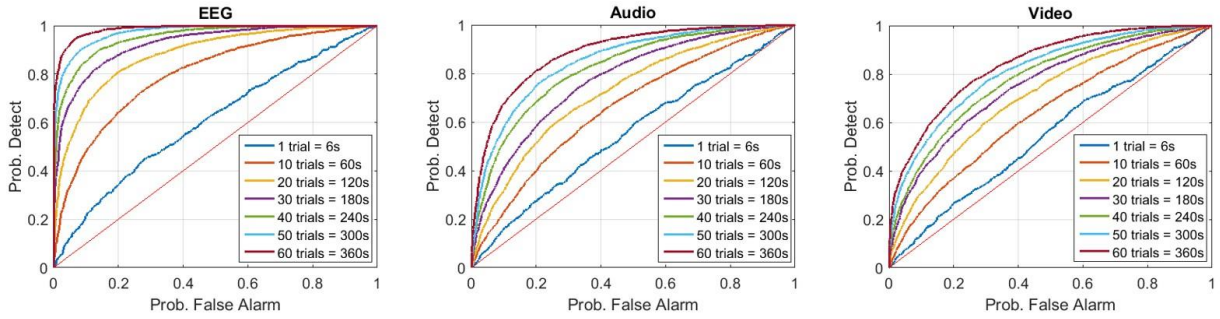

Figure 1. Probability of detection versus false alarm for EEG (*left*), audio (*middle*), and video (*right*) modalities. Each panel provides ROCs as a function of increasing number of trials from 1 to 360, corresponding to 6 seconds (s) to 360 s (6 minutes) for low and high cognitive loads.

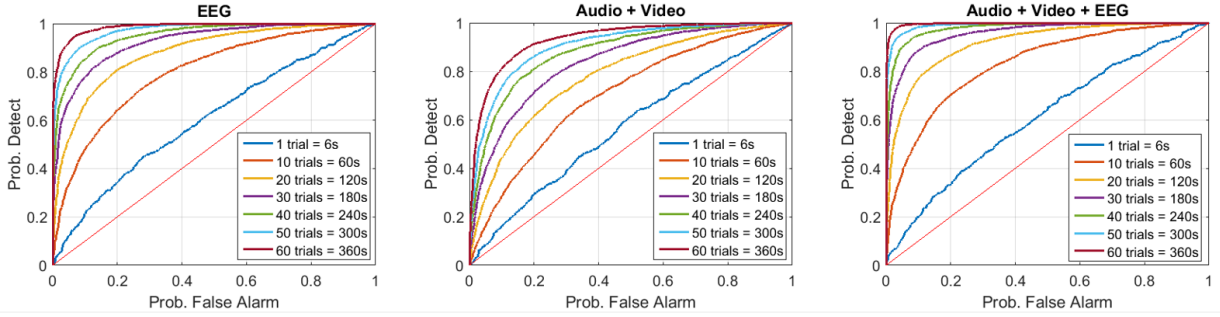

Figure 2. Probability of detection versus false alarm for combinations of EEG (*left*), audio (*middle*), and video (*right*) modalities. Each panel gives ROCs as a function of increasing

number of trials from 1 to 60, corresponding to 6 s to 360 s (6 minutes) for low and high cognitive loads.

In other work, we have directly assessed the possibility of using speech-based features to estimate fatigue/performance levels, as measured using PVT. These preliminary experiments involved gathering speech and PVT data from one subject, three times daily over the course of one working week. After extracting speech features from the acoustic speech signal, those features were used to estimate their mean reaction time from the PVT via a multivariate linear regression model trained with leave-one-out cross-validation. Figure 3 shows the results of these prediction experiments. Estimates were accurate to 10 milliseconds (ms), even on this very limited and preliminary data set, which highlights the promise of speech features to act as a minimally invasive alternative to the PVT.

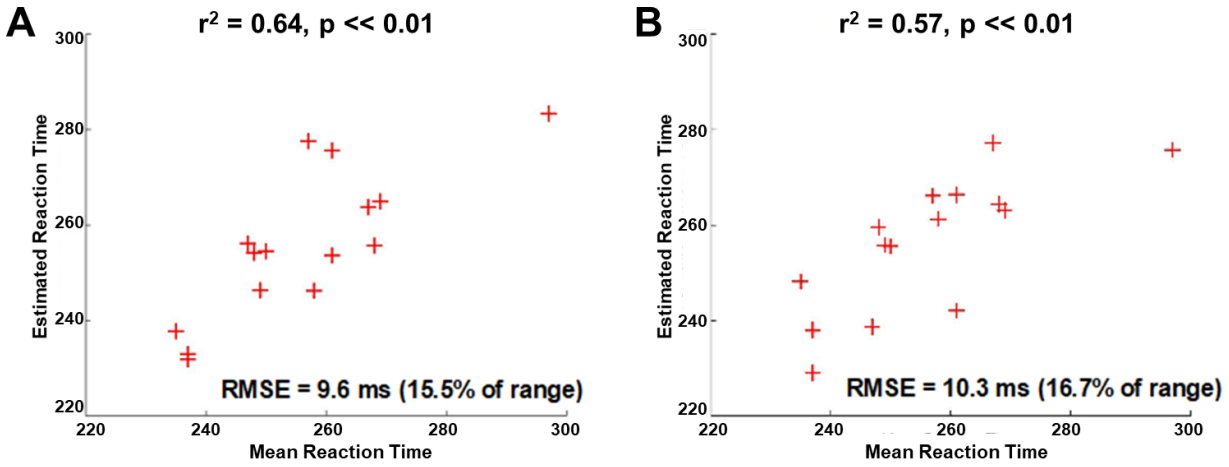

Figure 3. True mean reaction time, as measured using the PVT, versus reaction time estimated from speech features. (A) Left plot shows estimates obtained using features extracted from a conventional acoustic microphone. (B) Right plot shows estimates obtained with features from an inverse-filtered version of the same speech recordings, meant to imitate the output of a contact microphone placed on the neck.

Simulation and virtual reality create a natural testbed and key step toward a field-ready system. For the dismounted Warfighter, we used an immersive virtual reality environment that simulated aspects marksmanship, physical load (movement), and cognitive load (working memory) (Rao et al., 2020). This study found that both gait and speech features were highly predictive of the cognitive load state during this task. We also were able to predict the performance of both a working memory task and marksmanship using only physiological features.

Multimodal Sensor Measurements in Speech Processing

MIT LL's early work in multi-sensor analysis of speech and voice introduced an important paradigm in merging signals from acoustic and non-acoustic sensors when high levels of acoustic noise are present in the acoustic environment. Noisy environments in this early work included Black Hawk helicopters, high mobility multipurpose wheeled vehicle (HMMWV)

tanks, and urban warfare. Advances in non-acoustic sensors, including skin vibration, bone conduction, accelerometer, and microwave radar sensors, provide the exciting possibility of both glottal excitation and, more generally, vocal tract measurements that are relatively immune to acoustic disturbances and can supplement the acoustic speech waveform. Moreover, non-acoustic sensors have the ability to reveal certain speech attributes lost in the noisy acoustic signal; for example, low-energy consonant voice bars, nasality, and glottalized excitation. We introduced a novel framework and led a research effort for combining the output of these sensors according to their capability in representing specific speech characteristics in different frequency bands for use in a variety of applications including low-rate speech encoding (Quatieri et al., 2006) speech enhancement (Tardelli et al., 2003), and automatic speaker authentication (Campbell et al., 2003). We began with an empirical development of this multi-sensor framework (Quatieri et al., 2006), later followed by a first-of-its-kind theoretical "optimal" framework (Tardelli et al., 2003) with metrics that rely on frequency-dependent signal quality and signal-to-noise ratio.

Methods

Our primary goal is identification of behavioral biomarkers for monitoring fatigue and performance in operational situations. Achieving this outcome requires three types of data: (1) gold standard assessment of fatigue, (2) flight performance data, and (3) neurophysiological measurements. Reaction-time data, gathered using the well-established Psychomotor Vigilance Test, constitutes a behavioral metric that is the current gold standard for measuring effectiveness, fatigue, and alertness. Actual flight performance data provides the bridge between changes in behavior in the laboratory and the potential impact on mission performance. Speech (acoustic and non-acoustic) recordings constitute behavioral data that reveals neurocognitive function through changes in motor control. Additional modalities provide complementary information about neurocognitive function, and provide contextual information regarding physical effort, fatigue, and arousal; they will be collected but not analyzed as part of the current scope. Therefore, these three sources of data will together form the basis for accomplishing our primary goal, with potential complementary contributions from physiological measures (e.g., electrodermal arousal, heart rate variability, ocular movements).

Objective

The objective of the cognitive fatigue assessment protocol is to create an operationally relevant task such that measurements of voice, physiology, and key flight metrics may be assessed. We discuss our development process, which culminated in our final experimental protocol that was administered. We also discuss the complete experimental timeline and process to demonstrate how we adapted to human subjects research during the COVID-19 global pandemic.

Development Process

To ensure we had a clean experimental design that would have the potential of meeting our objective of an operationally relevant flight scenario, we iterated and consulted with several domain experts.

We consulted with two instructors at the Bedford Aeroclub at Hanscom Airfield, MA. From them, we were introduced helicopter ground school basic knowledge, an opportunity to experience a helicopter flight simulator, and advice on metrics of performance. We used the helicopter simulator experience there as a touchstone for constructing our own simulator at MIT LL. Further, we used the helicopter manual for the Robinson R44 helicopter as a way of relating levels of performance to general aviation standards for privately licensed and commercially licensed helicopter pilots. Therefore, we could put deviations in performance from our flight protocol into context.

We also consulted with an Army helicopter pilot from the U.S. Army Aeromedical Research Laboratory (USAARL), a U.S. Air Force fixed-wing instructor pilot, and a helicopter pilot at MIT LL. All three individuals shaped the makeup of the final task as well as the type of speech that would be collected.

Closed Loop Pattern

The final, core task we asked participants to perform consisted of closed loop traffic patterns around Hanscom Air Field in virtual reality for nominally 90 minutes of continuous flying.

Figure 4 shows the traffic pattern loop performed by participants as well as the corresponding speech points. The loop starts at position 0. Before takeoff from the ground, participants first speak their intent: “Hanscom Tower, Bravo 314. Current speed and altitude are 110 feet (ft), 0 knots. The time is twelve forty, ascending to 500 feet, seventy five knots.” Participants then are asked to ascend and travel to the corner of runway 11, which is the first corner of the square. Then, the participant proceeds to the second corner at point 1, and crucially should be upon arrival at the target altitude and speed of 500 feet, 75 knots. The participant is instructed to maintain this speech and altitude as closely as possible for the straightaway portion between points 1 and 2 in the figure. During this downwind leg of the loop, the participant is asked to speak a second time: “Hanscom Tower, Bravo 314. Current speed and altitude are XX and YY. Will begin descent from base leg.”

At point 2 on Figure 4, the participant is directed to begin descent with the intent of cornering the final runway end point and then landing either on the takeoff area, point 0, or on the triangular region and oriented along the runway. The additional instruction to land at the triangle region was added partway through collection as a way of providing an explicit visual cue for just where touchdown was expected. Upon landing, or crashing, the participant spoke a third time: “Hanscom Tower, Bravo 314. Successful or unsuccessful landing. Current speed and altitude are 0 and 110 knots.”

This space is intentionally blank.

Figure 4. Traffic pattern protocol executed by participants for nominally 90 continuous minutes of flight.

Upon landing or crashing, the participant used a virtual reality (VR) reset button to position himself once more in the takeoff location and the process was repeated until nominally 90 minutes had elapsed. The entire loop is approximately 447 meters and took approximately 3 to 5 minutes to complete one repetition. The 90-minute duration was informed by consultation with our pilot experts as a typical sortie duration.

We had originally constructed more specific grading guidance for the ascending and descending portion of the traffic pattern following the guidance from the R44 manual. However, the majority of our participants were novice pilots and additional specifics would not have been feasible for them to perform. We also had contemplated a high stress emergency maneuver to end the protocol in order to study the dynamic change between a low load condition and a sudden high load condition. While we experimented with ideas of inducing a forced auto-rotation or requiring travel to another unfamiliar airport for an emergency landing, the additional complexity ultimately was not appropriate at this stage of the research and the skill level of our participants.

Experimental Timeline

Because this research program had to be executed during a global pandemic, we took steps to protect the safety of the participants and the experiment proctors. One way in which we minimized human-to-human proximity was to perform the informed consent process virtually. All participants gave written informed consent in this MIT Institutional Review Board-approved study. The informed consent process was completed using the DocuSign (www.docusign.com) service. Consenting was done prior to arrival for the experimental session.

Figure 5 shows the timeline of a participant's involvement over the study. The informed consent process was performed prior to the experiment day when possible. However, participants had to perform a self-attestation report 24 hours prior to the study to attest to being COVID-19 negative in fact and symptom. Participants began by performing two minutes of speech reading

and free speaking via image response. This served as a potential baseline voice recording. Then, participants were instrumented with physiological sensors, which took approximately one hour. Next, metadata surveys regarding general health and fitness as well as actual and virtual flight experience were completed. Then, participants were given the opportunity to free fly in virtual reality in order to acclimate to the experience. Once participants demonstrated that they could perform the desired traffic pattern, the participant performed the second speech collection outside of VR, the Psychomotor Vigilance Test, and the self-report Stanford Sleepiness Scale. Next, a five-minute physiological baseline was collected, in which the participant sat quietly with eyes closed. At this point, the participant was given final instructions for the 90-minute traffic pattern and flew nominally continuously for 90 minutes. Upon conclusion of the 90 minutes of flight, a third speech collection was performed, followed by a second PVT and a second physiological baseline.

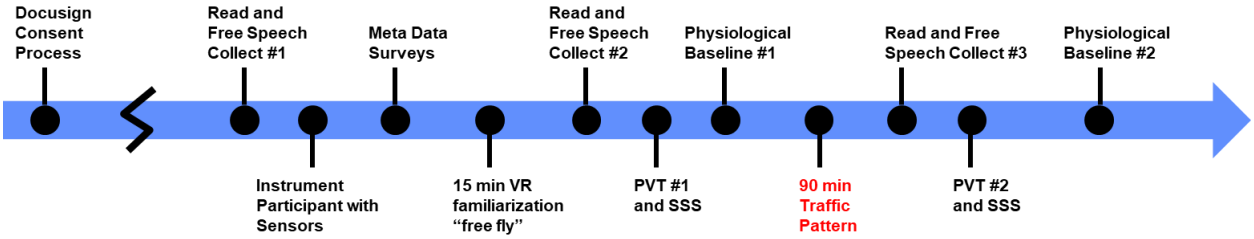

Figure 5. Participant experimental timeline beginning with a remote informed consent process prior to the day of the study. Participants completed a battery of neurophysiological assessments in addition to the 90-minute traffic pattern flight.

The three speech collections, two prior to the 90-min flight, and one following the 90-min flight, were additional to the speech that was collected throughout the flight. Though not part of the planned analysis, they were collected as snapshot points for assessing gross changes in cognitive state. Similarly, the PVT assessment acts as a potential gold standard for changes in cognitive state. The Stanford Sleepiness Scale is a complementary subjective assessment of alertness to the objective measures of speech and voice.

Multimodal Signal Collection

The strength of a highly multimodal setup, like the one used here, is that it enables a higher sensitivity of detecting changes in cognitive state and a higher specificity of characterizing the nature of cognitive change. In this experiment, we collected a range of physiological, behavioral, and contextual signals, which, when taken together, may provide a deep understanding of the subjects' cognitive state.

Lab Streaming Layer.

Lab Streaming Layer (LSL) is a system for the unified collection of time series data in research experiments (https://github.com/sccn/labstreaminglayer). The application coordinates communication between data collection modalities as well as time synchronization and data storage. At the heart of the LSL application is a graphical interface called Lab Recorder (LR). LR receives data from all the modalities and stores the data into a standardized extensible data format, or "xdf" format. To enable modalities to stream data to LR, an LSL-specific bridge needs

to be written. Through the course of this program, MIT LL wrote several of these bridges to enable modalities to be streamed into LSL. As a whole, the LSL package includes the LR aggregator and the set of modality-specific bridges that stream the data to LR.

LSL is an extensible, cross-platform, open-source framework. At present, MIT LL has added many sensors and modalities to the LSL framework. However, there is no limit to the number of sensors that can be added in the future. One might imagine adding new devices to test and validate against existing sets of devices, or based on the experimental requirements, selecting from a list of possible modalities to record.

One motivating factor for choosing the LSL framework was that it was designed with real-time processing in mind. In the future, algorithms developed offline, such as those proposed in this report could be implemented to provide real-time feedback to the user for applications such as cognitive fatigue detection.

Physiological Sensor Selection.

In this study, we primarily used Shimmer Sensing wearable sensors to measure a range of physiological signals (http://www.shimmersensing.com/). Though other systems were evaluated, the Shimmer platform was selected because it all met our criteria for integration. The selection criteria included:

- Wireless Streaming: The system needed to stream data (Bluetooth or WiFi) so a custom bridge to LSL can be written. The presence of an application programming interface (API) or software development kit (SDK) was crucial to building that LSL bridge.

- Mobility: Subjects would need to move their arms, legs, and head while flying the helicopter simulator in virtual reality. Therefore, a wearable sensor that allowed for a range of body movement was preferred.

- Long Duration Recording: The system needed to record data over many hours as the experimental plan, including the setup and miscellaneous time, required a system that may need to record data for up to six hours or longer.

- Secure Digital (SD) Card Logging: Though not critical, the capability for simultaneous data logging to an SD card was preferred as a redundant measure in case there were streaming interruptions.

- Multimodality: Rather than build LSL bridges for many different sensor systems, a single platform capable of performing multimodal data collects was preferred.

The Shimmer Sensing platform met these criteria as the platform has wireless streaming, is fully wearable, can log and stream data for over 10 hours (or more), and is capable of capturing any bio-potential signal source.

Alternate evaluated sensor selections.

Before settling on the Shimmer Sensing platform, several other platforms were evaluated.

To measure the electrocardiogram (ECG), the Bittium Faros was tested (https://www.bittium.com/ medical/bittium-faros). The Faros is regarded as a model device for its high data quality. Further, it met the requirements stated above. MIT LL also wrote a LSL bridge for the system to stream data into LSL. However, upon pilot testing the system, the Bluetooth connectivity seemed inconsistent. Across Faros devices, there was variability in whether the devices would connect to the computer Bluetooth. Given the lack of reliability of the Bluetooth connectivity, which was an important criterion, the Faros was abandoned as a potential device.

To measure electrodermal arousal (EDA), the wrist-worn Empatica E4 device was tested (https://www. empatica.com/research/e4/). Just as with the Faros, the Empatica E4 seemed to meet the criteria and an LSL bridge was written for the device. However, upon running pilot studies with the device, MIT LL's review of the data during test flights showed that the data quality was rather poor. If the subject did not move their hand at all, then the data quality was suitable. However, if there was hand movement, the data quality dropped dramatically. Naturally, in an experiment where near constant hand movement is expected to fly the helicopter, this device was deemed not suitable.

The Biopac system (https://www.biopac.com/) is a high quality, multimodal data collection platform. The main drawback of the system is that it is wired and requires the subject to be tethered to a power source. While this might have worked in the current setup, a tethered connection is not in the envisioned future concept of operations. Further, the Shimmer system was more affordable than the Biopac per setup, which allowed development to proceed in parallel at multiple remote locations.

In addition to LSL, MIT LL also evaluated the program 'Microsoft Platform for Situated Intelligence' (PSI; https://github.com/microsoft/psi). PSI, like LSL, is an open, extensible framework for research of multimodal systems. Though potentially more powerful than LSL, the overhead associated with developing bridges to PSI was much more than with LSL. For ease of use, LSL was selected over PSI.

Virtual Flight Simulation

Flight Simulator.

The flight simulator chosen for this experiment was the X-Plane 11 simulator (https://www.x-plane.com/). In addition to being a high fidelity simulator in its own right, the X-Plane 11 software had several key features that made it ideally suited for this program. First, it has native capability to integrate virtual reality headsets, such as the HTC Vive Pro Eye. Second, there already exists a robust mechanism to stream X-Plane 11 data into LSL using the XPlaneConnect bridge (https://github.com/nasa/XPlaneConnect). Third, the data saving process, native to X-Plane 11, is easy to use and served as an extra source of redundancy on the flight behavioral data. Taken together, the X-Plane 11 software was the ideal choice for the data collection.

Helicopter Model.

The simulated helicopter model used in this experiment is the Robinson R44 Raven II.

The model was chosen as it is considered one of the easier helicopters to fly. Many learn to fly using that model and as such, the applicability to the helicopter pilot community would be broad. The specific model used within the X-Plane 11 software is the R44 2.0.0 model, updated for X-Plane 11 v11.33 (https://forums.x-plane. org/index.php?/files/file/52056-robinson-r44-raven-ii/). The model used in this experiment was free to download and use. However, expert helicopter pilots commented that the model did not fly as realistically as they would have liked. Specifically, it seemed that the virtual helicopter fell to the ground much faster than expected when the thrust vector was reduced. There may be other, newer models of helicopters that exist that may be more realistic in the virtual environment (e.g., https://store.x-plane.org/Robinson-R44-Raven-II_p_1315.html).

Virtual Reality Headset.

The HTC Vive Pro Eye headset was used in this data collection. There are a number of potential options to select from when picking a VR headset with eye tracking enabled. The HTC Vive Pro Eye was selected for two reasons. First, the MIT LL team has used Tobii Eye Trackers extensively in the past and trust the data quality. The eye tracking system embedded in HTC Vive Pro Eye headsets are Tobii systems and therefore, the data quality would be reliable. Second, there already exists an SDK for the integration of the HTC Vive headset. MIT LL wrote a custom LSL bridge for this module to capture both eye tracking and pupillometry data (Smalt et al., 2021). Other headsets may be selected, but one should note that a new LSL bridge would need to be developed.

Summary of Sensed Signals

Table 1 summarizes the selected sensors and systems and Figure 6 shows a subset on a participant. The five Shimmer devices used are listed based on their location on body and the signal of interest. The two factors influencing the decision on sampling rate were the signal of interest and the capability to stream at that rate over Bluetooth. Note that for Shimmer devices where a higher sampling rate was required, only the signal of interest was recorded/streamed (e.g., ECG). On devices with lower sampling rates, multiple signals were recorded/streamed on the same device (e.g., respiration and accelerometry).

Figure 7 provides an example set of signals derived from the sensors in in Table 1. This visual is just a short section of the total 90-minute traffic pattern, wherein the subject performs five loops. These signals show the raw data and therefore best represent the signal quality from each sensor.

This space is intentionally blank.

Table 1. Summary of Sensors and Sensed Time Series Modalities.

Signal	**Sensor/System**	**Location on Body**	**Sampling Rate** (Hertz [Hz])
Electrocardiogram	Shimmer 1	Torso	512 Hz
Respiration	Shimmer 2	Torso	128 Hz
Accelerometry	Shimmer 2	Torso	128 Hz
Electromyography	Shimmer 3	Right Forearm	512 Hz
Electrodermal Activity	Shimmer 4	Right Hand	128 Hz
Accelerometry	Shimmer 4	Right Hand	128 Hz
Electrodermal Activity	Shimmer 5	Left Hand	128 Hz
Accelerometry	Shimmer 5	Left Hand	128 Hz
Eye Movements	HTC Vive Pro Eye	Head	250 Hz
Pupillometry	HTC Vive Pro Eye	Head	250 Hz
Speech	Webcam Mic	-	44.1 KHz
Voice	MIT LL Collar Mic	Neck	44.1 KHz
Electroencephalography	EasyCap + Smarting Amp	Head	256 Hz
Helicopter Controls	Puma Pro Flight Trainer	-	100 Hz
Helicopter Behavior	X-Plane 11	-	20 Hz

Note: The numbers associated with the Shimmers correspond to unique devices. Multiple signals (e.g., respiration and accelerometry) may be recorded on the same device.

This space is intentionally blank.

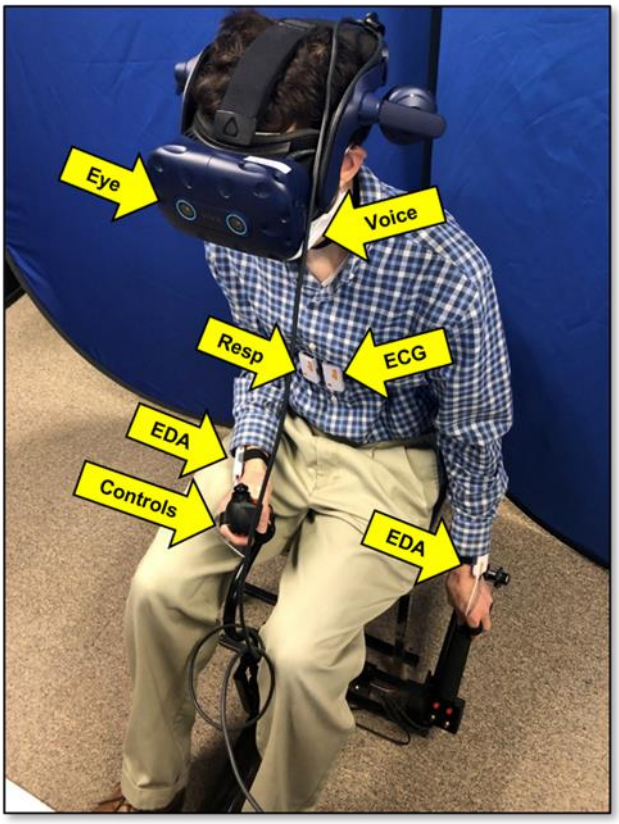

Figure 6. Image of several physiological sensors and behavioral systems on a participant. Not visible in the picture are the electromyography (EMG) sensor on the right forearm, the EEG cap (not used by this participant), and the Webcam microphone collecting speech data.

This space is intentionally blank.

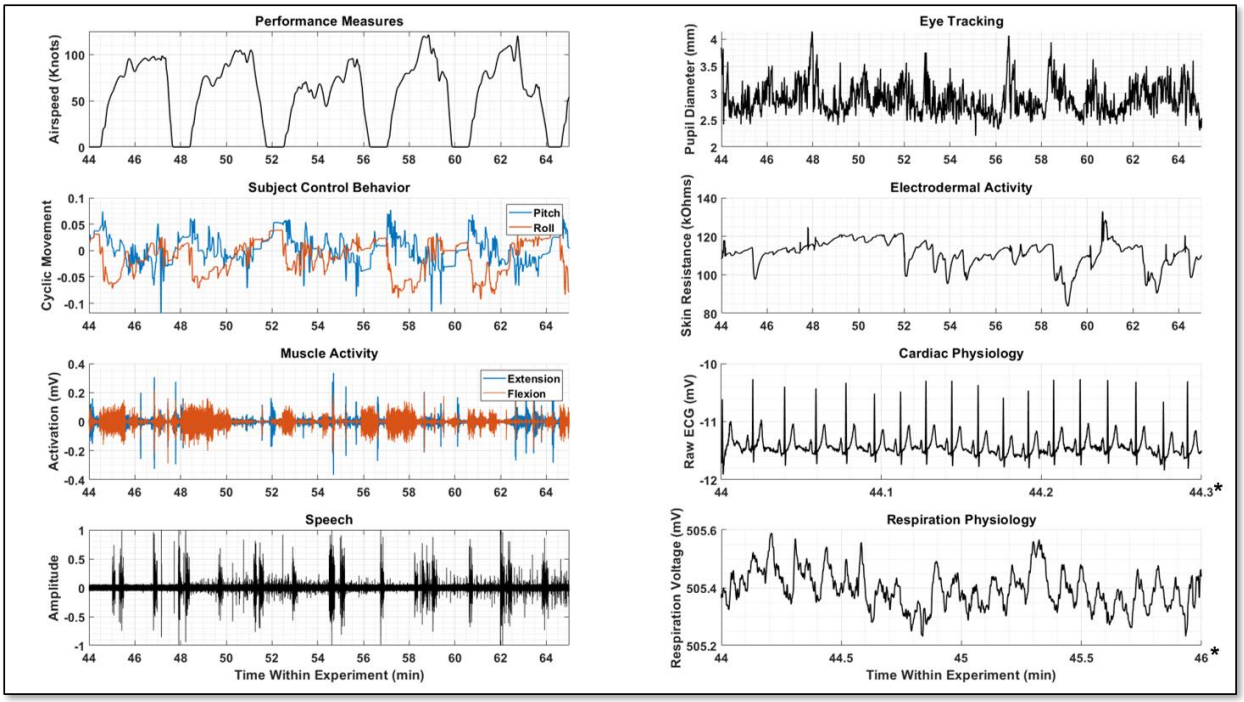

Figure 7. Representation of raw multimodal physiological and behavioral time series signals. The ECG and Respiration plots have different time axes to better visualize the data.

Additional Measurements

In addition to the primary measurements of acoustic speech, flight performance, and general physiology, we collected several other data streams of interest.

Psychomotor Vigilance Test.

Figure 8 shows a picture of the PVT box used to measure alertness. Alertness is assessed by rapidity of response to a stimulus. PVT has been shown to be both sensitive and repeatable to cognitive fatigue and has been used extensively in sleep research in particular. We developed custom, micro-controller-based PVT boxes for the administration of a three-minute PVT before and after the flight. The interstimulus interval was uniformly distributed between one and four seconds.

Stanford Sleepiness Score.

The Stanford Sleepiness Score is a one question, 7-level self-report score of mental alertness. (See appendix for a copy). A score of 1 corresponds to “Feeling active, vital, alert, or wide awake” and a score of 7 corresponds to “No longer fighting sleep, sleep onset soon; having dream-like thoughts.” This score was assessed just prior to the 90-minute traffic pattern and just after with the intent of observing if there were self-perceived changes in cognitive state.

Wireless Voice Monitor.

The wireless voice monitor (WVM) is an accelerometer that captures a noise robust

signal generated from speaking by virtue of being placed above the collarbone and below the thyroid prominence (also known as the Adam's apple). This signal is more robust to ambient noise than a regular microphone, and the signal is approximately confidential in that only the voicing act of speech is captured (much like a humming sound) rather than actual words. This sensor was developed by MIT LL and piloted in several participants in this study as an exploratory sensor. We summarize its capabilities and performance compared to a regular microphone here, and refer to its publications for more details (Mehta et al., 2017; Chwalek et al., 2018; Mehta et al., 2019).

Figure 8. Psychomotor Vigilance Test box. Participants performed a three-minute reaction time test in which they pressed the right button with the right index finger in response to the blue light turning on. A green light signaled the end of the test.

The WVM streams data from two on-board sensors: a high-bandwidth analog accelerometer (frequency response: 0 - 5 kilohertz [kHz]) and a micro-electromechanical system MEMS acoustic microphone (frequency response: 100 Hz - 15 kHz). We sometimes refer to the accelerometer as a contact microphone, but in practice, an accelerometer is more robust to acoustic noise than a contact microphone due to the accelerometer being a surface vibration transducer. The signal for each sensor is saved with sampling rate of 44.1 kHz and bit depth of 16 bits.

Table 2 compares a set of speech and voice features that are commonly extracted from a regular microphone and denotes when that feature can also be extracted from the WVM. The WVM, as expected, can extract the voice related features of fundamental frequency, harmonic to noise ratio, creaky voice quality, and cepstral peak prominence. However, by the WVM's nature, the WVM does not capture formants or mel frequency cepstral coefficients as these two features are dependent upon the upper vocal tract for shaping the voiced air flow into speech.

Electroencephalography Sensor.

Electroencephalography (EEG), while not a primary sensor to be analyzed for this study, was collected on two willing participants. Though not analyzed in this report, the data is available for exploration into more direct measures of neural function than speech and other biomarkers. While EEG still has many practical hurdles before it could be field-worthy, it has shown promise for detecting brain state, and may be of interest.

Table 2. Speech Features Available from Wireless Voice Monitor (WVM) and Regular Microphone

Speech Feature	Definition	Use	WVM	Microphone
Fundamental Frequency	Vibration rate of vocal folds	Prosody	X	X
Harmonic to Noise Ratio	A ratio of periodic and aperiodic speech components	Voice quality	X	X
Creak	Phonation with incomplete glottal closure	Voice quality	X	X
Cepstral Peak Prominence	Relative height of f_0 peak in cepstrum above linear trend line	Voice quality	X	X
Formants	Resonances of vocal tract	Articulation		X
Mel Frequency Cepstral Coefficients	Discrete cosine transform of the log of the magnitude of the Fourier transform	Articulation		X

Analysis Methods

In this section, we report our analysis methodology for the various collected data streams.

Psychomotor Vigilance Test.

The PVT is a sustained attention and reaction time test in which a participant must monitor a LED light and press a button as quickly as possible as soon as the light illuminates. We conducted a PVT assessment before and after the 90-minute traffic loop in order to determine with a gold standard metric the change in cognitive fatigue.

We report the median reaction time and the 25th and 75th percentiles for the reaction times, as well as the difference in the medians, post- minus pre-. We used Mood's median test to determine for each participant if the change in median reaction time was significant.

We used robust statistics in the form of percentiles to analyze the data because of a sensor acquisition hardware fault in which occasionally a button press by a participant was not properly registered. Missed button presses lead to abnormally large reaction times. While it is possible that these reaction times are true attentional lapses, based on follow up interviews with participants, clearly hardware did play a role. Consequently, we cannot definitely rule out the data points and instead rely on the majority of the responses being correct and representative of behavior.

Table 3. R44 Traffic Pattern Performance Standards (source: R44 Flight Training Guide, March 2019)

Metric	Private Standard	Commercial Standard
Airspeed	10 kts	5 kts
Elevation	100 ft	50 ft

Flight Analysis Methods.

Table 3 reproduces the key traffic pattern performance standards explicitly covered in the R44 manual. The R44 manual defines performance for two grades of pilot qualifications: a commercially rated pilot (the more stringent qualification), and a private pilot. Therefore, the assessment reduces to measures of accuracy and stability at attempting to maintain a desired position and orientation in space. Consequently, we created metrics of variability that captured these two ideas that summarize performance on a single complete traffic loop.

We computed measures of variability using the L1 and normalized L1 signal length metric (i.e., the sum of the absolute value of the first order difference of the signal). The L1 metric was appropriate because the scenario was dynamic with changes in elevation and orientation naturally as part of the loop. A single scalar mean which would be needed for standard deviation as a variance metric was not applicable.

The formula for path length is:

$$L_1 = \sum_{n=0}^{N-1} |x[n+1] - x[n]| \tag{1}$$

The normalized path length is:

$$L_{1,norm} = \frac{1}{N} \sum_{n=0}^{N-1} |x[n+1] - x[n]| \tag{2}$$

Specifically, the signals in our L1 analysis metrics were groundspeed, elevation above ground level, heading, pitch, and roll. We additionally computed the duration of a loop and the projected loop distance onto the Earth (i.e., how far did the helicopter travel if it had been on the ground the whole time).

To determine start and end points of the loop, the elevation and groundspeed were first visualized. Then, the point of first ascent was manually identified for each loop as the loop start point and the point at which ground speed returned to zero was used as the loop end point.

Speech Prosodic Analysis

Speech prosody is the timing and intonation of speech. Prosody is often termed the "melody" of speech, and has been shown to be a sensitive indicator of cognitive state. Pitch, which is the human perception of fundamental frequency, is a central carrier of prosodic information. We focused on the pitch waveform as a time series wherein statistical changes may be indicative of cognitive state.

Figure 9 and Table 4 show our speech prosody feature extraction pipeline and additional details on the final feature computations from the extracted pitch waveform. We used Praat (https://www.fon.hum.uva.nl/praat/), an open source software application, to extract the fundamental frequency and perform voice activity detection. For females, we set the pitch ceiling at 300 Hz and for males, we set the pitch ceiling as 275 Hz. These values were determined based on inspection of the raw waveforms.

For the speech variability metrics, we computed summary statistics (e.g., mean, standard deviation, skew, kurtosis, and percentiles) on each vocalization prior to takeoff. Specifically, we have only focused on the speech prior to takeoff because we are interested in the predictive nature of speech biomarkers for the end goal of averting safety incidents or mission compromise. Nonetheless, the in-flight speech and the post-landing speech could be similarly analyzed, and their information might be combined to provide a better estimate of cognitive state on the current loop in order to improve prediction on the following loop. This strategy would be as opposed to, or in addition to, using the speech immediately prior to takeoff on that subsequent loop.

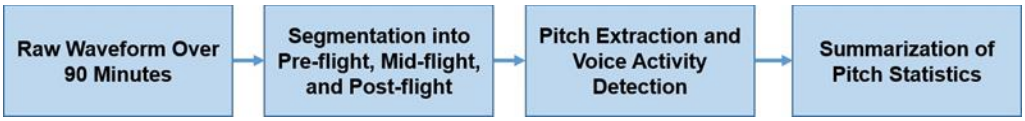

Figure 9. Speech prosody feature extraction pipeline in which pitch is extracted from segmented waveforms and statistically summarized.

Table 4. Speech Summary Statistics

Metric	Definition
Mean	Unbiased sample mean
Median	Middle value of the ordered set
Mode	Most frequent value determined from a histogram of 30 bins
Std	Unbiased sample standard deviation Range
Range	computed as max minus min
IQR	Interquartile range computed as $75^{th} - 25^{th}$ percentile
Mid90	Range computed as $95^{th} - 5^{th}$ percentile
Skew	Measure of the asymmetry of a distribution
Kurtosis	Measure of the prevalence of values far from the mean
Min	Minimum value
Max	Maximum value
Percentiles	5 to 95 in steps of 5%

Joint Flight and Speech Analysis

We conducted a joint flight and speech correlation analysis to get at the central question of this research: is there a useful relationship between flight performance and speech biomarkers? Specifically, we pairwise correlated each combination of the 30 speech statistics and 12 flight performance metrics using Spearman correlations across the trials within a participant. Before performing the subject level correlation, a robust outlier detection method was applied to screen out traffic loops with potentially erroneous data artifacts. We computed the interquartile range (IQR) on the metric of interest, and then any points that fell more than 1.5 times the IQR above the 75th percentile or below the 25th percentile were excluded.

To move beyond correlations to predictions, we constructed a binary classification framework. Using only the speech collected immediately prior to takeoff, we attempted to classify whether the subsequent traffic pattern would be a high variability or low variability loop for the participant.

Each participant's set of loops and features were zero-meaned and standardized to a standard deviation of one. Then, we used a leave one subject out cross validation framework in which we predicted each of the valid traffic loops for the held out test participant.

The $n - 1$ participants are used to identify a single median value, which is nominally zero, to separate the training data into high variability and low variability traffic loops. For the held out test participant, a separate median threshold is computed to separate that participant's loops into high and low variability. The median (versus the mean) is used in order to force a nearly perfect, class-balanced, binary label set.

The $n - 1$ participant's data is then used to construct m logistic regression classifiers where each classifier uses only a single speech feature. The m classifiers each predict a high or low variability score between 0 and 1 for each traffic loop of the test participant. The m predictions are averaged to reduce variability. The final integrated prediction for each test traffic loop is used to construct a receiver operating characteristic area under the curve.

Unlike with random forest classifiers, logistic regression is essentially deterministic, so variability is not a concern for performance prediction on the held out participant, i.e., the standard deviation of three repeated iterations is expected to be zero.

Results

Dataset Description

Collected Data.

This pilot experiment was a successful demonstration of a complex, multimodal signal acquisition system and multi-phased experimental protocol, all of which was developed and deployed during a pandemic. Despite these challenges, we collected data on six participants for the main experiment and a seventh for a technology demonstration. While there were occasional technical issues, a core dataset has been established for analysis within this report as well as

future exploratory analyses. A complete description of what is available is provided with the dataset itself.

Participant Demographics.

This dataset was collected with appropriate caution and safety measures during the 2020 COVID-19 global pandemic. We successfully recruited seven participants in spite of these challenge circumstances. Each provided written informed consent to participate in this MIT Institutional Review Board-approved study. While we had originally intended to recruit primarily experienced rotorcraft pilots, our eligible pool of participants consisted of a mix of novices, fixed-wing, and rotary-wing pilots.

Table 5 shows the flight experience level of the participants. Two participants were complete novices with no actual or virtual reality flight experience in fixed or rotorcraft. However, they were each permitted to practice flying for approximately 60 minutes prior to the 90-minute flight to mitigate learning effects that might confound with fatigue changes during the 90-minute flight. The other two novices had no real fixed wing or rotorcraft experience but did have some experience with both types in simulation. These individuals were given a brief familiarization period prior to the actual data collection. Of the remaining two participants, one was an experienced fixed-wing pilot and the other was an experienced rotorcraft and fixed-wing pilot.

Table 5. Participant Flight Experience Demographics

ID	Fixed-Wing Pilot	Rotorcraft Pilot
200	1	0
202	1	1
204	0	0
206	0	0
415	0	0
530	0	0

In terms of demographics, two participants were female and the rest male. All ethnically self-identified as white except for one who identified as Indian. The median age was 37 years old, the minimum was 31, and the maximum was 61. The limited gender and ethnic diversity should be kept in mind when attempting to apply these results to other populations. Follow on work should increase the diversity of the participant pool in order to mitigate any potential gender or ethnic biases present in this relatively homogeneous sample.

Measures of Cognitive Fatigue

The primary aim of this research is the investigation of whether and how physiology, specifically the voice, correlates with objective measures of flight performance. To situate this investigative question in the broader current understanding of cognitive fatigue, we collected two complementary measures of cognitive state.

Psychomotor Vigilance Scores.

Figure 10 and Table 6 summarize the results of the PVT pre- and post-assessments. For four of the six participants, the median reaction time increased, and for the remaining two, the median reaction time decreased. This change was significant for participants 415 and 530 at $p < 0.05$.

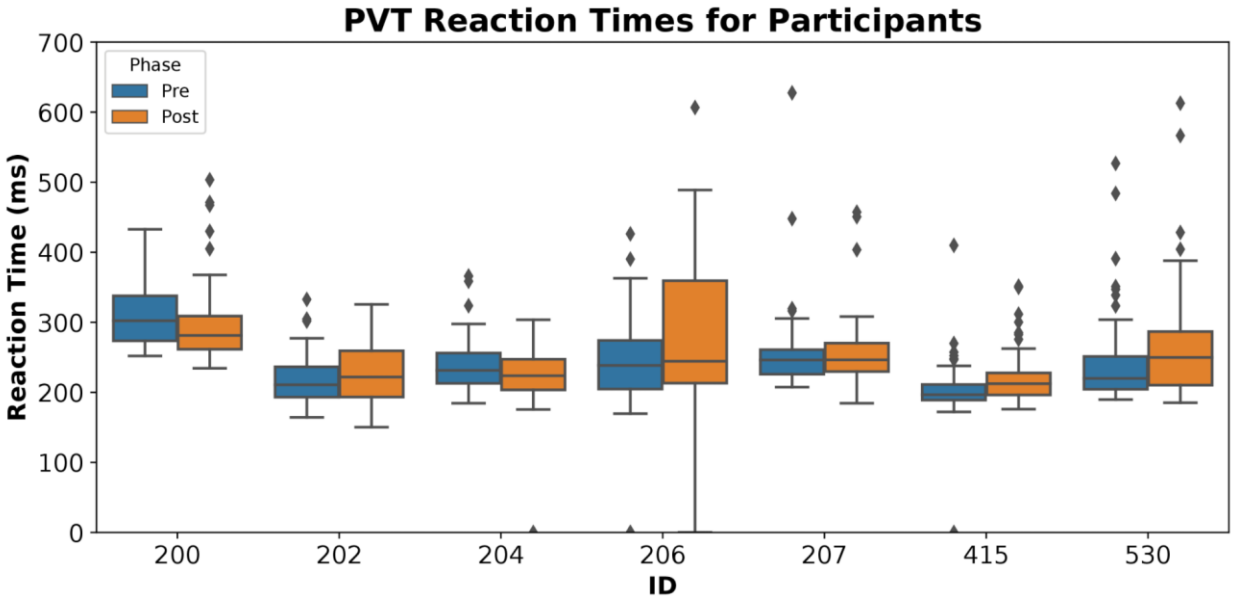

Figure 10. Participant PVT reaction times. Times greater than 700 ms are not shown for clarity.

Table 6. Psychomotor Vigilance Test Performance

ID	Session	Pre Reaction Time ms 50^{th} (25^{th} - 75^{th})	Post Reaction Time ms 50^{th} (25^{th} - 75^{th})	Prcnt. Change %
200	20201214	302.29 (273.41 - 338.63)	281.28 (261.35 - 309.04)	-6.95
202	20210112	210.75 (193.13 - 237.65)	221.87 (192.84 - 259.43)	5.28
204	20210203	231.33 (212.62 - 257.68)	223.75 (203.33 - 247.36)	-3.27
206	20210308	238.38 (204.32 - 274.23)	244.36 (212.84 - 360.23)	2.51
415	20201208	196.65 (188.96 - 211.29)	212.46 (196.31 - 228.64)	8.04
530	20201201	219.86 (204.24 - 251.85)	249.87 (210.26 - 287.31)	13.65

Figure 11 shows an example set of reaction times over time for one participant pre- and post-90-minute flight. Consistent with the majority of participants, this participant does show an increase in median reaction time. There are also potential hardware fault outliers at trials 1 and 31 (false positive) in the pre-flight session.

The PVT provides an independent baseline level of assessment on how much, if any, cognitive state changed over the course of the experiment. While a significant change was only noted for two participants, four of the six participants did show an increase in median PVT reaction time. Consequently, when we con- sider the flight behavior and speech dynamics we

may have reason to believe that decrements in performance or changes in speech dynamics may also be present.

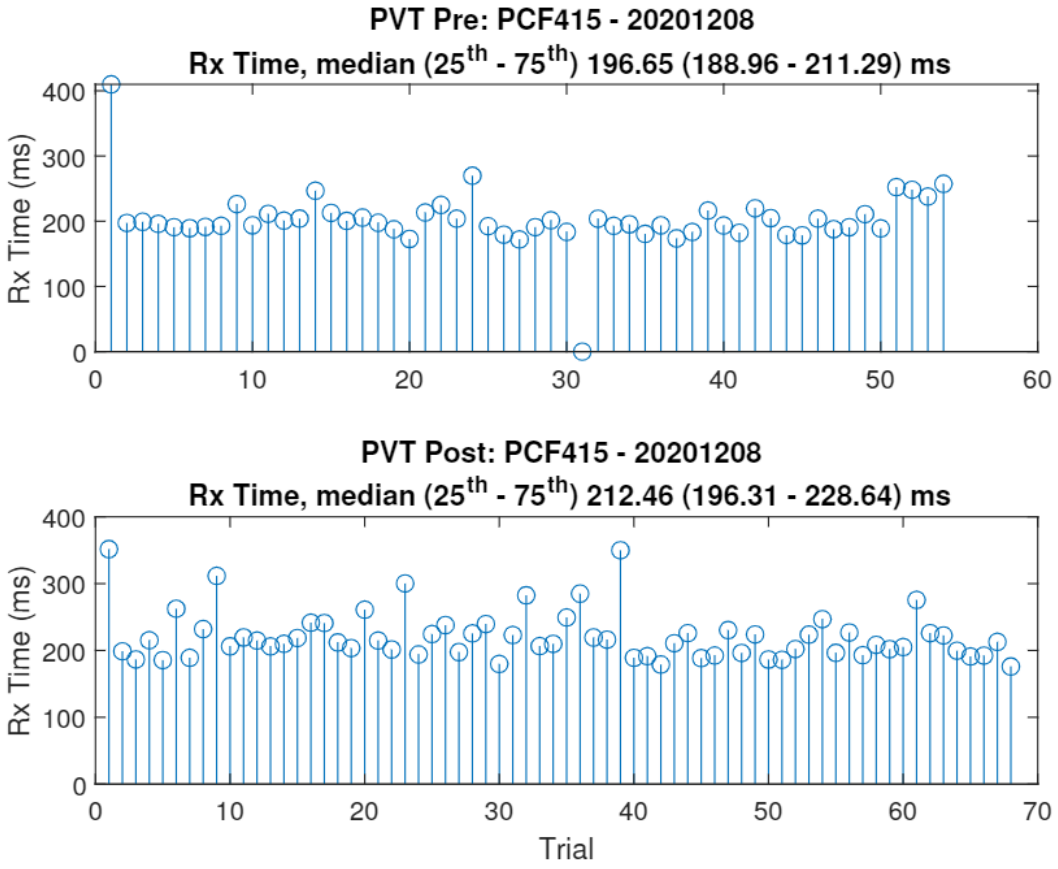

Figure 11. Pre-flight (*top*) and post-flight (*bottom*) Psychomotor Vigilance Test reaction times. Trials 1 and 31 in pre-flight data may be hardware faults. The overall median response time increases from pre- to post-, signaling the potential for an increase in cognitive fatigue.

Stanford Sleepiness Scores.

Table 7 summarize the results of the Stanford Sleepiness Score (SSS). Three of the six participants reported a decrease in alertness (increase in score) by at least one level, two reported no change (of these two, one was already considerably tired), and one reported an increase in alertness after the study. Overall, the results suggest qualitatively that participants generally perceived the flight task to be mildly cognitively fatiguing or at least not stimulating.

Table 7. Stanford Sleepiness Scores: Pre- and Post-Flight

ID	Session	Pre-Flight	Post-Flight	Change
200	20201214	2	3	1
202	20210104	1	1	0
204	20210115	2	3	1
206	20210308	3	2	-1
415	20201208	6	6	0
530	20201201	2	5	3

Flight Performance Over Time

We first introduce examples of collected flight data over time at the time series level for a single participant to provide a grounding in the raw flight data. We then apply our performance variability metrics to this participant and all the participants to provide a picture of the spectrum of intrapersonal and interpersonal performance ranges. The amount of within subject performance variability is important for subsequent analysis, as this within subject variability is ultimately what we will attempt to predict with physiological measurements.

This space is intentionally blank.

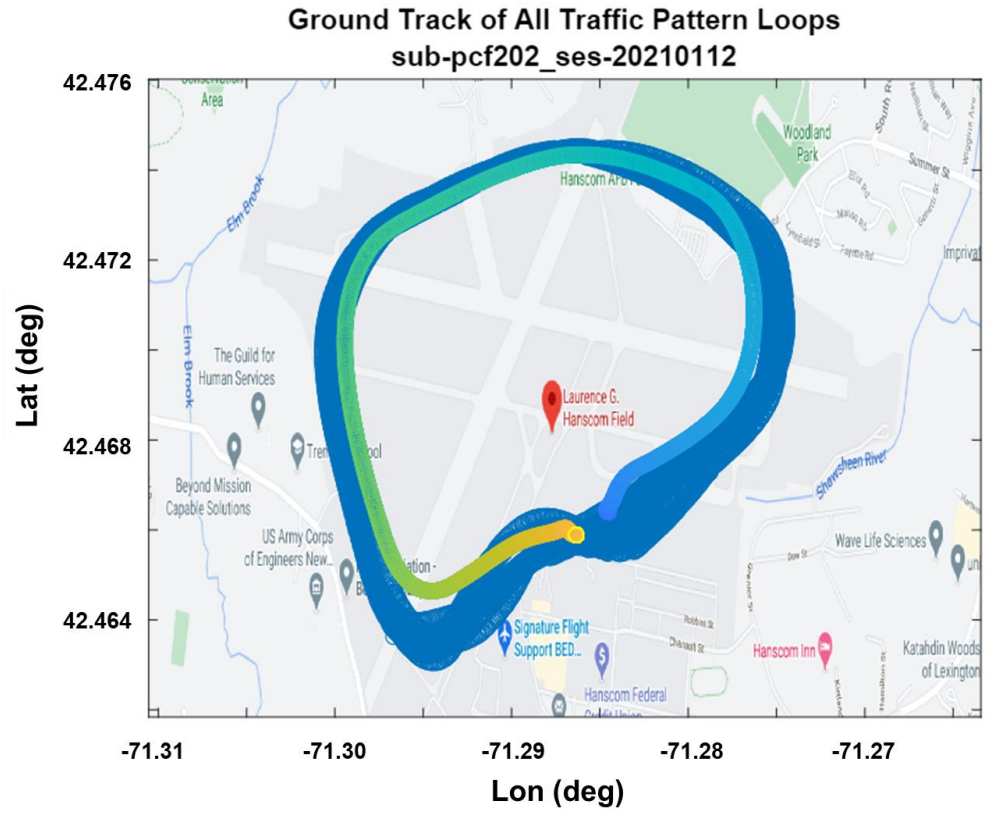

Figure 12. Expert flight performance: ground track pattern over Hanscom Air Force Base. All traffic loops are superimposed (blue), and a single loop is highlighted and color-coded from start (blue) to finish (yellow). Map Data copyright 2021 Google.

Flight Performance Time Series.

Figure 13 shows the traffic pattern ground track overlaid on Google Maps' view of Hanscom Air Force Base. This figure shows all the loops overlaid on top of each other, and the small amount of dispersion is indicative of the expert handling of the helicopter. A single loop is color coded by time from takeoff (blue) to touch down adjacent to the takeoff point (yellow).

Figure 13 and Figure 14 show several other performance time series monitored during the traffic pattern. The elevation and speed over time are relatively similar from loop-to-loop, and the orientation likewise is relatively uniform loop-to-loop. Because the traffic pattern is a closed loop, we expect the heading to progress through three hundred sixty degrees of change, which is what is seen. We also would not expect large, random deviations in roll and pitch but only changes that would be relevant for banking during a turn and ascending or descending. Again, the patterned nature of the orientation changes suggests the helicopter is being adequately controlled.

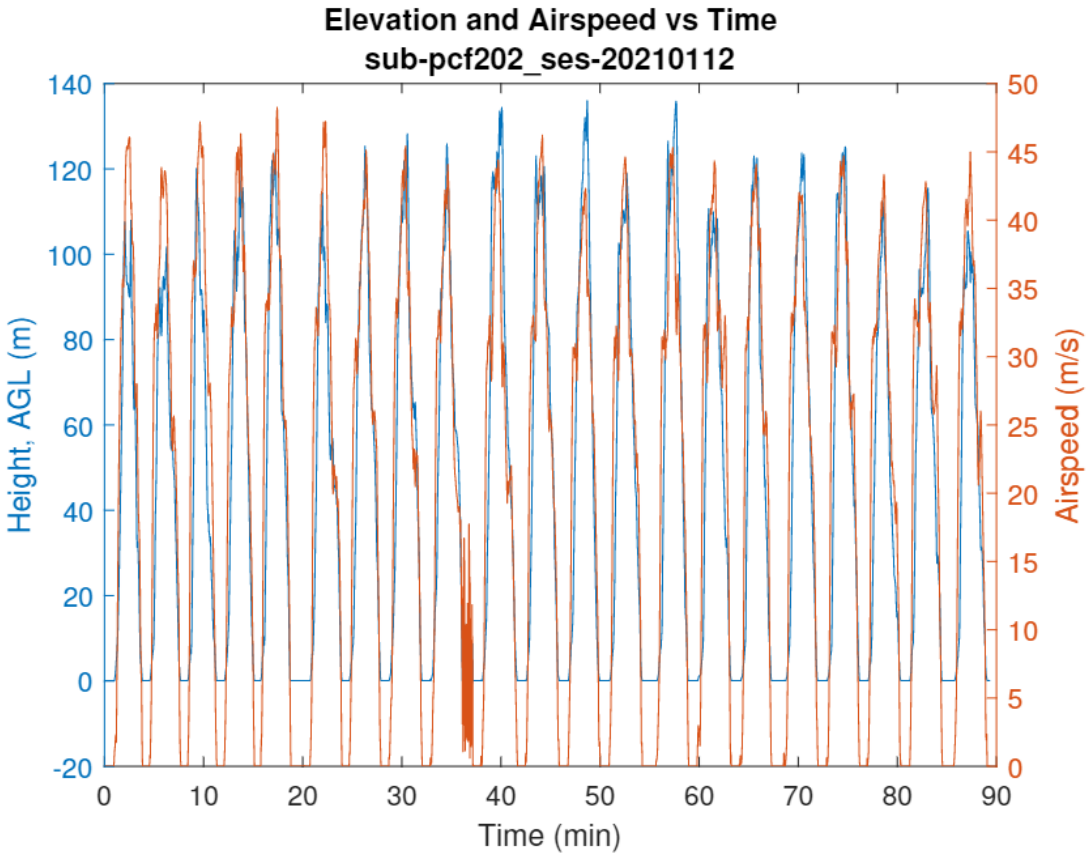

Figure 13. Expert flight performance: elevation and airspeed follow a highly stereotyped pattern indicative of strong loop-to-loop consistency.

Flight Performance Variability.

Figure 15 provides a snapshot of the range of inter-participant and intra-participant variability for each of the flight performance metrics. For each metric, across all loops for all participants, the minimum value of the metric was used as a normalization value. Therefore, a plotted value of "two" for example means that the participant's value for that loop was twice the minimum value among all the analyzed loops. This normalization scheme allows interpreting all the metrics as a multiple of the minimum. Because all of the L1 metrics are measures of variability, lower is generally considered better performance.

The variability among participants is highly indicative of the amount of reported flight experience. By far, the least variable participant across the metrics was the expert pilot. The next least variable participant was the actual fixed-wing pilot, then the two helicopter VR-experienced novices, and finally, the two novices who had no prior experience other than an extended training session. We had to clip the y-axis of the plot in order to show the expert's variability values, otherwise they would be too compressed. That such a strong correlation exists among flight experience and measures of variability is excellent support for these measures capturing meaningful aspects of flight performance.

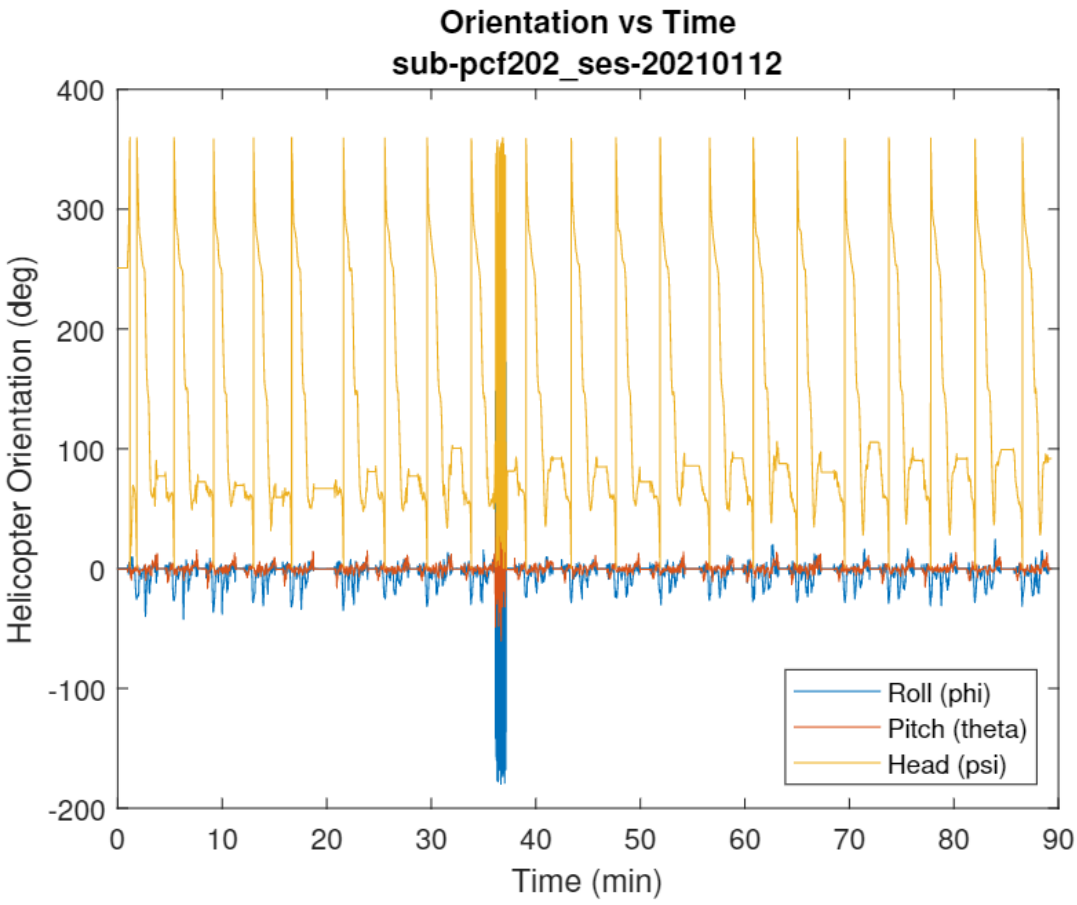

Figure 14. Expert flight performance: Helicopter rotational orientations of heading, pitch, and roll follow a highly stereotyped pattern indicative of strong loop-to-loop consistency.

This space is intentionally blank.

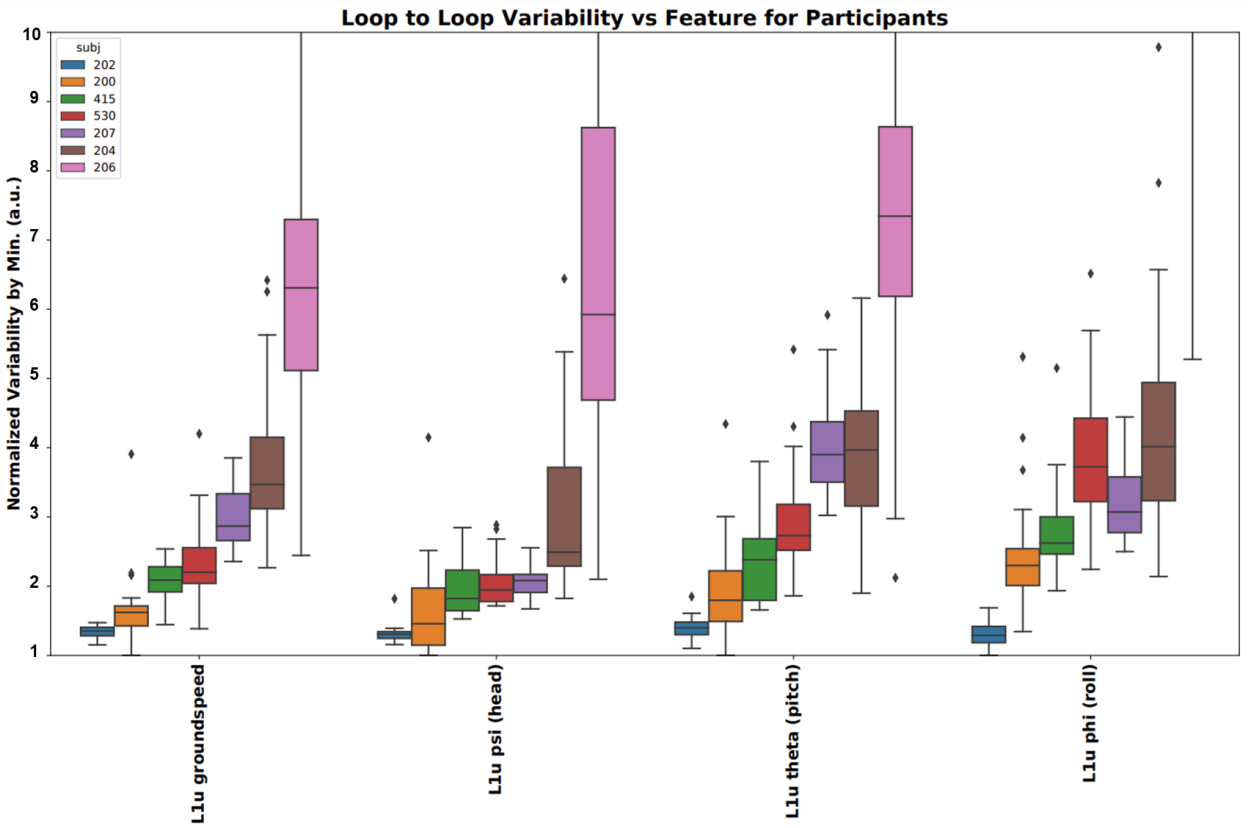

Figure 15. Normalized loop flight performance variability. The expert pilot clearly stands out as having the most control (least variability), and the other participants' variability correlates with degree of experience.

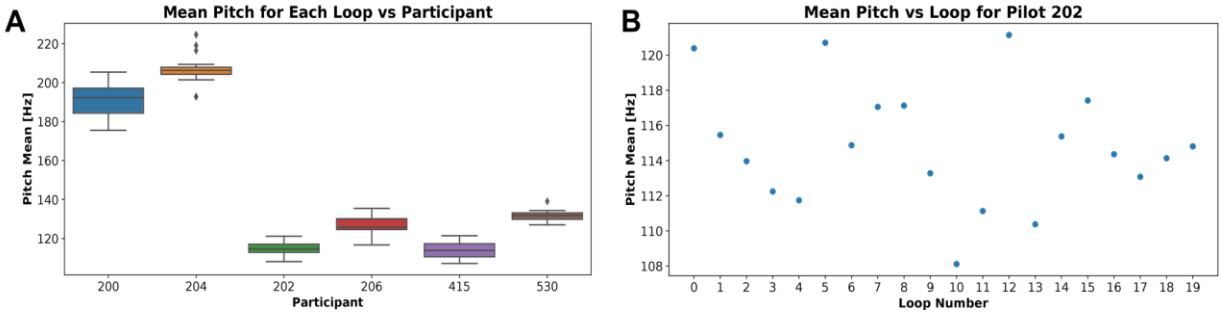

Figure 16. (A) All participants: Variability in loop-to-loop mean pitch. (B) Expert pilot: loop-to-loop mean pitch.

Vocal Biomarkers over Time.

Figure 16 provides a high-level summary of one of the speech prosody biomarkers, mean pitch over an utterance. Figure 16a shows the mean pitch versus participant, and there is a clear separation between the females (participant numbers 200 and 204) and males. As expected, the females have higher overall pitch values on average. Of particular interest is not the absolute value in the pitch but the variability in pitch from loop to loop. Therefore, we normalized within

participant by the minimum pitch to quantify the percentage change from the minimum. The median percent change in median pitch among all participants was 6.5%.

Figure 16b shows the mean pitch for the expert pilot versus loop iteration. For this participant, there is no obvious trend in change in pitch with respect to time. However, this might be expected as over a relatively short period of 90 minutes, there may be moment-to-moment changes in concentration than can dominant any general trend. Therefore, this figure serves as motivation for the next analysis in which we attempt to correlate loop-to-loop variability in speech with loop-to-loop variability in performance.

Joint Flight Performance and Voice Analysis

Correlation of Performance with Voice.

Figure 17 shows the correlation between speech prosody biomarkers and flight performance for the expert pilot. In Figure 17a, we threshold the correlations to have a magnitude greater than 0.3 and an uncorrected p value less than or equal to 0.05. In Figure 17b, we show as a scatter plot an individual flight variability metric vs speech biomarker. Each point in the plot represents the performance-speech pair for one traffic pattern loop. Across all the loops, there is a clear positive correlation in which more variability in the elevation is associated with more variability in the pitch (Spearman correlation of 0.59). In other words, variability in motor control can be seen in both flight control and also speech control.

Figure 18 shows the results of applying this framework to all the participants. Figure 18a shows the number of occurrences in which a flight variability-speech biomarker pair meets the magnitude threshold of 0.3 and the p value criterion of less than or equal to 0.05. Because we are not requiring a consistent sign for this count, it is possible that two participants may react strongly but opposite from each other in terms of speech-flight behavior. Figure 18b shows the scatter plot for the speech pitch mean and the variability of the helicopter pitch feature pair. This correlation is significant for two participants.

This space is intentionally blank.

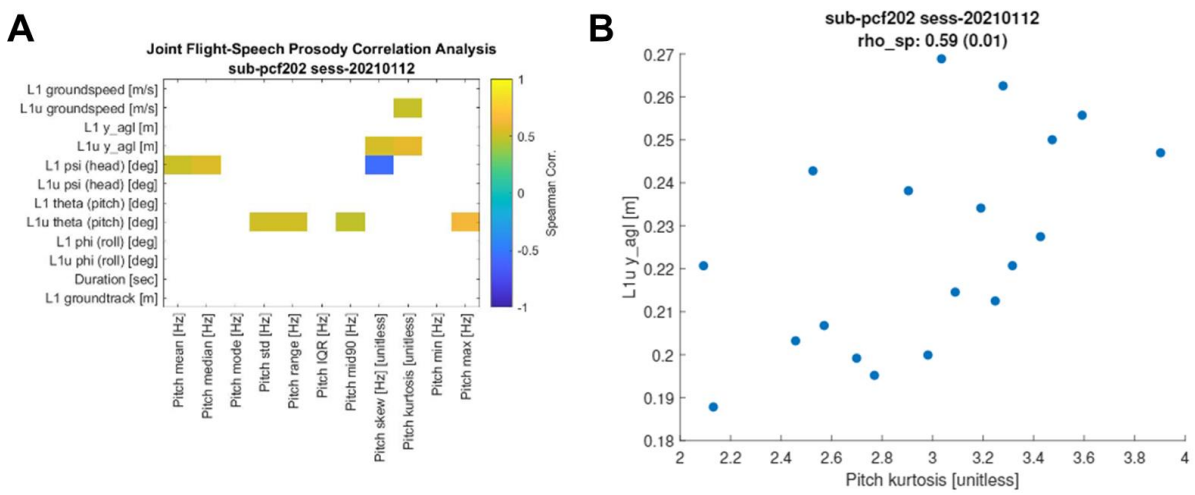

Figure 17. (A) Expert pilot, all speech-flight correlations. (B) Scatter plot for a specific feature pair.

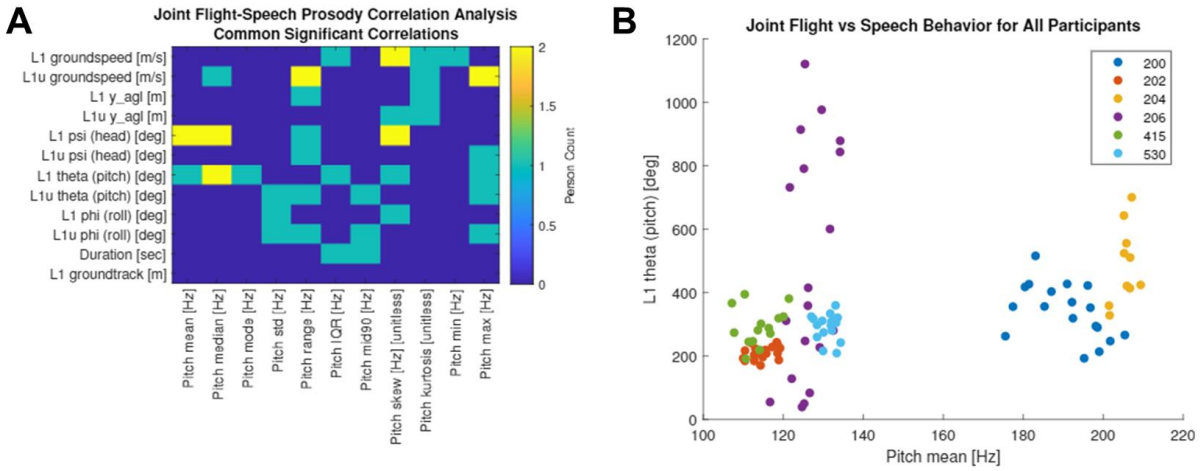

Figure 18. Joint correlation analysis. (A) Count of the frequency of significant correlations. (B) Scatter of correlations for all participants.

This space is intentionally blank.

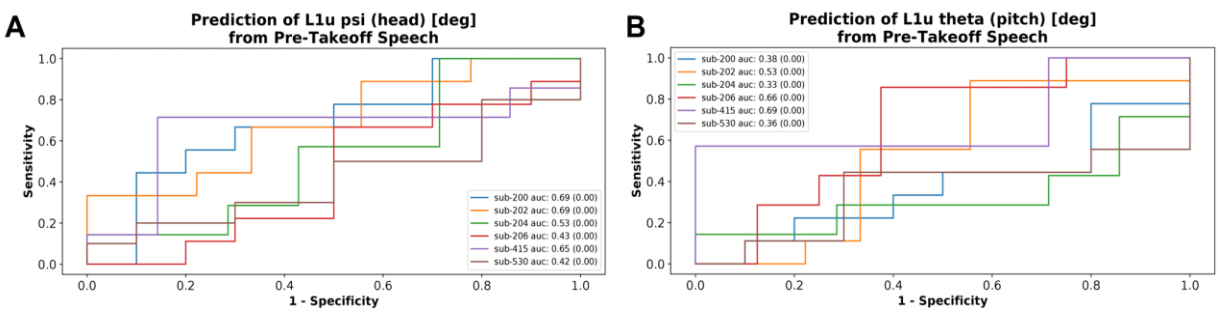

Figure 19. Prediction of high vs low variability loops from pre-takeoff speech for participants. (A) ROC for predicting heading variability. (B) ROC for predicting pitch variability.

Prediction of Flight Variability with Pre-takeoff Speech.

Figure 19 shows the classification prediction results of high vs low variability loops using pre-takeoff speech in receiver operative characteristics (ROC). Across all the variability metrics, variability in heading was most consistently predictable across the individuals (Figure 19a). With other metrics (e.g., Figure 19b), the results are closer to chance (0.50).

One possible reason for heading being particularly predictive is that it may be most sensitive to the experience level of the pilot, and therefore provides the cleanest signal for how stable an individual's motor performance may be. The three ROC curves with the largest AUC correspond to some of the most experienced pilots: the actual expert helicopter pilot, the real fixed-wing pilot, and an experienced VR pilot. Pitch and roll variability actually have natural visual feedback in the form of a horizon line whereas heading requires a pilot to focus on and maintain a more ambiguously defined trajectory through space.

Acoustic and Non-acoustic Vocal Biomarkers in Noise

Wireless Voice Monitor.

Figure 20 shows an example demonstration of fundamental frequency extracted from the WVM accelerometer signal (ACC) and the regular microphone (MIC). There is excellent agreement visually between the time series for this segment of speech with a correlation of 1.0 (perfect).

As is typical in speech analysis, fundamental frequency doubling or halving errors can occur when processing either ACC or MIC signal. Even in moderately noisy environments, the average fundamental frequency can sometimes be extracted to a sufficient degree using the microphone signal. However, in high noise environments, faithful extraction of the fundamental frequency (e.g., for low-bitrate communication) is challenging. This was one of the main reasons throat-mounted sensors were applied in the 1960's by Cambridge Air Force Research Laboratory.

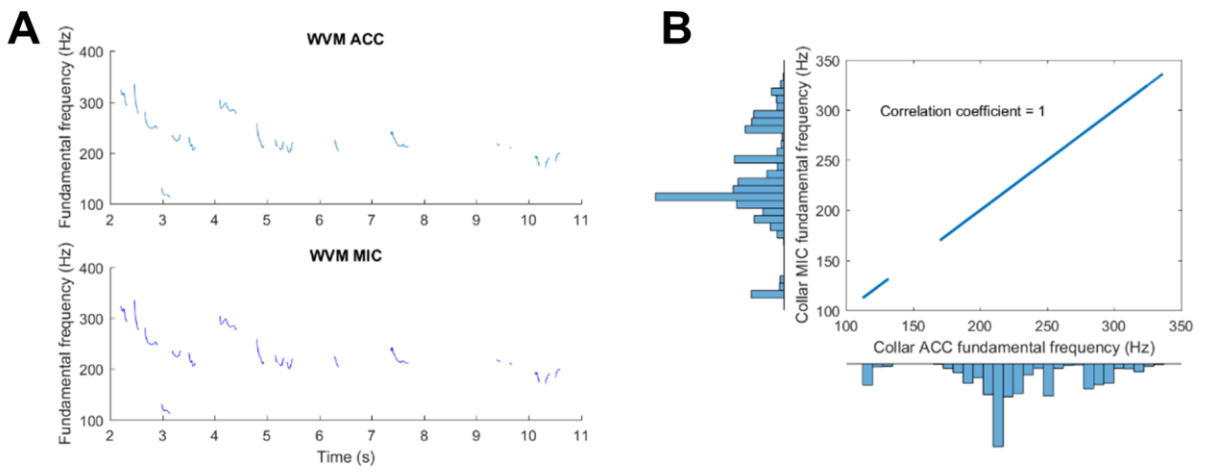

Figure 20. (A) Comparison of the fundamental frequency time series extraction from the WVM accelerometer signal (top) and the collocated regular microphone (bottom). Visually there is excellent agreement between the two waveforms. (B) A scatter plot of the fundamental frequency values from (A) and the marginal histograms. The high correlation of 1 quantifies the excellent agreement seen in (A).

Technology Demonstration in Noise.

One participant, separate from the analysis of those previously discussed, also completed the protocol as a technology demonstration of the wireless voice monitor in noise. Helicopter rotor noise generated by the X-Plane 11 R44 software model was played out through two Electro-Voice speakers (model QRX 212/75) which have the capability of 600 W continuous power handling and 2400 W peak and a single Electro-Voice model QRX218S subwoofer. The ambient measured sound level in dBA was approximately 88 dBA. For reference, without explicit noise injection, ambient levels are approximately 42.5 dBA in the room. Noise levels inside a real helicopter can exceed 100 dBA (Kupper et al., 2004).

Figure 21 shows amplitude versus time and spectrographic visualization of the wireless voice monitor's acoustic microphone and the non-acoustic accelerometer signal for a segment of speech. The acoustic microphone signal is severely degraded by the ambient helicopter noise whereas the non-acoustic signal is virtually unaffected.

Figure 22 shows a side-by-side comparison of the time series waveform collected from the wireless voice monitor accelerometer and the webcam microphone. The wireless webcam is extremely robust to the ambient noise. There is little if any indication that ambient noise levels are quite high. By contrast, the webcam audio was corrupted and was unsuitable for analysis.

This was a striking demonstration of how the wireless voice monitor is still able to extract data that can be used to potentially predict pilot performance despite being in highly adverse acoustic recording conditions.

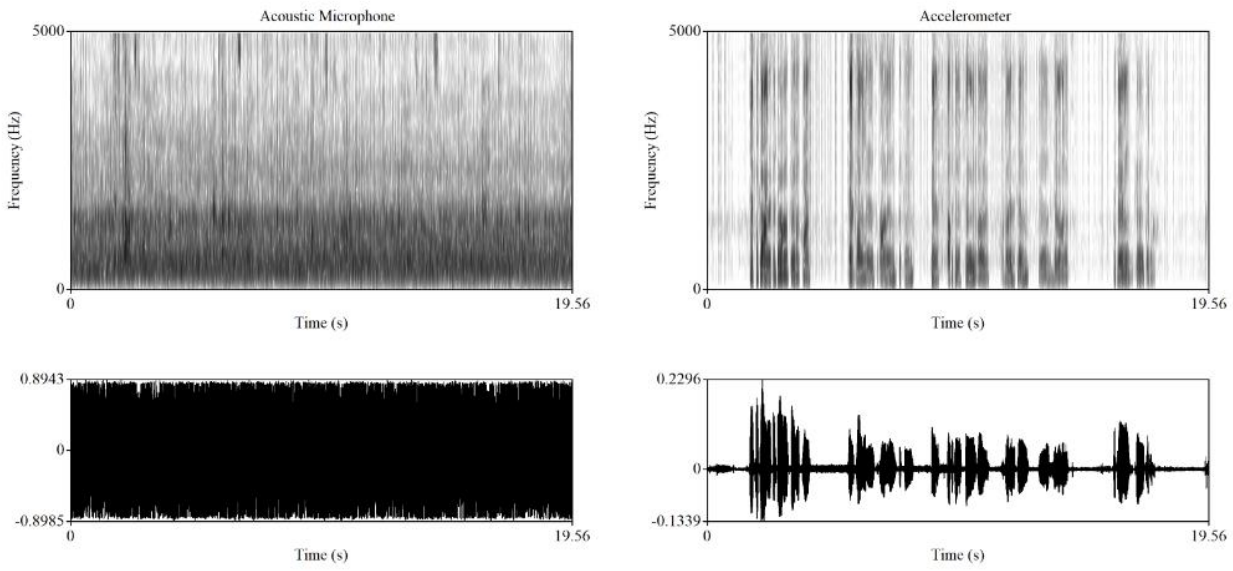

Figure 21. Amplitude versus time and spectrogram visualizations of the wireless voice monitor acoustic microphone (*left*) and noise robust accelerometer signal (*right*) dramatically contrast the signal degradation of the acoustic microphone in 88 dBA of ambient noise and the reliability of the accelerometer signal.

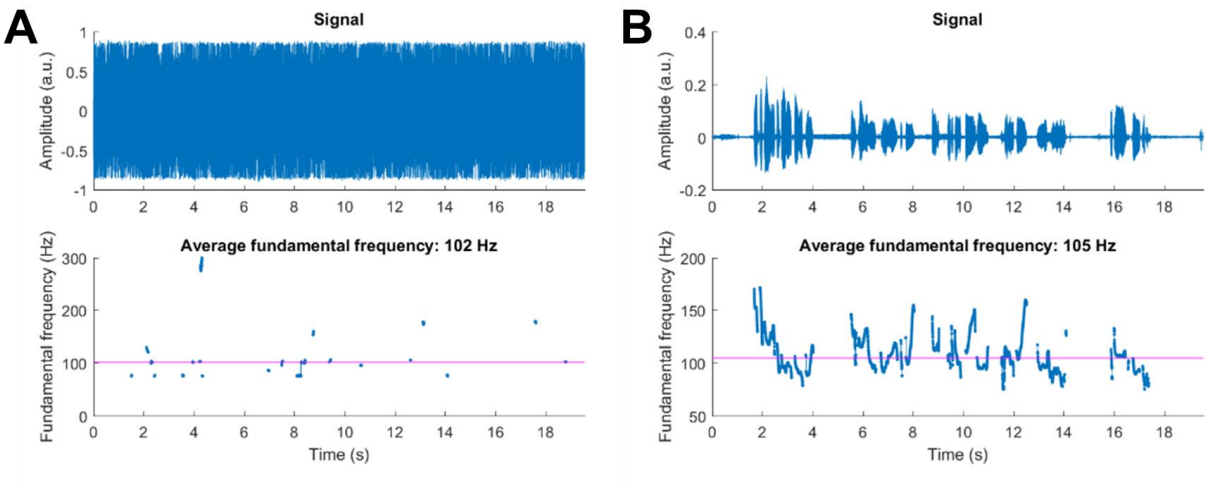

Figure 22. (A) Comparison of the fundamental frequency time series extraction from the WVM acoustic microphone versus (B) accelerometer signal. Fundamental frequency tracking completely fails with the acoustic signal whereas extraction is successful with the noise robust accelerometer signal.

Conclusion

The primary objective of this program was to develop and optimize non-invasive biomarkers (e.g., vocal) for quantitative, non-invasive measurements of pilot fatigue in operational environments. To satisfy this overarching objective, MIT LL was tasked with delivering:

1. A multimodal data collection of human participants during a cognitively fatiguing, operationally relevant, simulated task; and

2. A technical report quantifying predictive ability of vocal indicators of performance in the MIT LL flight simulator.

For the first deliverable, MIT LL has collected data from six participants during a multimodal flight simulator experiment that lasted nominally 90 minutes for each participant. Based on evaluation with the Psychomotor Vigilance Test, there was a trend for participants to exhibit increased reaction times after the traffic pattern exercise in the flight simulator and, by that definition, exhibited cognitive fatigue. For the second deliverable, MIT LL has analyzed and documented the analysis of the vocal biomarker element of data collection within this report.

Summary of Results

We saw a trend towards an increase in PVT reaction time following the traffic pattern protocol, which may be a sign of cognitive fatigue associated with the protocol. We also saw a clear trend for experience and flight variability, which strongly motivates using only experienced helicopter pilots in future studies.

With respect to the participants in this study, there was substantial within-participant variability both in terms of flight behavior and in speech biomarkers. Further, while participants often show a speech prosody and flight behavior coupling that is greater than 0.3 and has a p value less than 0.05, the exact feature pair varies among participants. That we do see a strong correlation for our expert pilot is a promising indicator that neuromotor coordination and fatigue has a common source that affects speech motor control and task relevant motor control.

We conclude that the 90-minute protocol is useful for studying declines in performance associated with cognitive fatigue. The study is best performed with expert pilots in order to avoid the confounds of inexperience and novelty that comes from novice pilots, and that in the expert pilots and others there is evidence of speech performance coupling, though the exact manifestations of these differences may be individual specific.

Limitations

The two primary limitations of this study have been the number of participants and the experience level of available participants. This study's small sample size of less than ten participants did suggest that vocal biomarkers may be sensitive to change in flight ability. However, we would strongly urge replicating against a larger population of at least thirty participants of diverse genders and ethnicities. Just as important as increasing the number of participants would be recruiting from an active duty military population whose occupational

specialty is rotorcraft flight. Involving this population, which was not possible in this pilot study because of distance and pandemic constraints, could provide further confidence that conclusions would be applicable to the population of interest.

Future Work

This initial study proved out the development and test scenario for a multimodal data collection platform. We see several logical extensions from this point. First, we would conduct a study using participants from the target population in a replica of our setup at a base that trains helicopter pilots and therefore has a ready pool of skilled individuals. Second, we would repeat the study's principal components inside a high-fidelity flight simulator such as a UH-60 Blackhawk simulator. Third, contingent upon continuing strong, promising results, we would work with a relevant base in order to obtain vocal recordings during actual flight operations. These vocal recordings, by their nature, may be too sensitive to capture in their entirety, and this motivates a related, fourth suggestion: further development of non-acoustic speech acquisition sensors that obtain speech like biomarker information without capturing intelligible speech.

Consistent with the overarching goals of this program, we believe the research done in monitoring cognitive state in Army helicopter pilots has a strong connection and relevance to dismounted Soldiers. Specifically, in future work, the move towards making sensors robust to operational noise exposure has relevance for both. Additionally, this sensing should ultimately be part of a closed-loop, real-time feedback system: one that is always running, but through the use of advanced artificial intelligence can manage operator workload seamlessly in a man-machine partnership and one which could act if the human pilot was in serious danger due to exhaustion. Non-invasive cognitive sensing will be the key enabler of future man-machine teams and safe, effective operations in the complex and strenuous operating conditions of tomorrow.

References

Campbell, W., Quatieri, T., Campbell, J., & Weinstein, C. (2003). *Multimodal speaker authentication using nonacoustic sensors.* Massachusetts Institute of Technology Lincoln Laboratory.

Chwalek, P. C., Mehta, D. D., Welsh, B., Wooten, C., Byrd, K., Forehlich, E., Maurer, D., Lacirignola, J., Quatieri, T. F., & Brattain, L. J. (2018). Lightweight, on-body, wireless system for ambulatory voice and ambient noise monitoring. *2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN), IEEE,* 205-209.

Darby, J. K., Simmons, N., & Berger, P. A. (1984). Speech and voice parameters of depression: A pilot study. *Journal of Communication Disorders, 17*(2), 75-85.

Fava, M., & Kendler, K. S. (2000). Major depressive disorder. *Neuron, 28*(2), 335-341.

Krishnan, A., Bidelman, G. M., Smalt, C. J., Ananthakrishnan, S., & Gandour, J. T. (2012). Relationship between brainstem, cortical and behavioral measures relevant to pitch salience in humans. *Neuropsychologia, 50*(12), 2849-2859.

Harnsberger, J. D., Wright, R., & Pisoni, D. B. (2008). A new method for eliciting three speaking styles in the laboratory. *Speech Communication, 50*(4), 323-336.

Horwitz, R., Quatieri, T. F., Helfer, B. S., Yu, B., Williamson, J. R., & Mundt, J. (2013). On the relative importance of vocal source, system and prosody in human depression. *IEEE International Conference on Body Sensor Networks*, 1-6.

Kupper, T.E., Steffgen, J., & Jansing, P. (2004). Noise exposure during alpine helicopter rescue operations. *Annals of occupational hygiene, 48*(5), 475-481.

Le, P. N., Ambikairajah, E., Choi, E. H., & Epps, J. (2009). A non-uniform subband approach to speech-based cognitive load classification. *7th International Conference on Information, Communications, and Signal Processing*, 1-5.

Levitt, H. (1971). Transformed up-down methods of psychoacoustics. *The Journal of the Acoustical Society of America, 49*(2B), 467-477.

Makyska, N., Quatieri, T. F., & Sturim, D. (2005). Automatic dysphonia recognition using biologically-inspired amplitude modulation features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, I-873.

Mehta, D. D., Chwalek, P. C., Quatieri, T. F., & Brattain, L. J. (2017). Wireless neck-surface accelerometer and microphone on flex circuit with application to noise-robust monitoring of lombard speech. *Interspeech*, 684-688.

Mehta, D., Deshpante, R., Letter, L., Froehlich, E., Siegel, A., Quatieri, T., & Brattain, L. (2019). On-body monitoring of voice-based cognitive load features in an auditory working memory task. *16th International Conference on Wearable and Implantable Body Sensor Networks (BSN), IEEE,* 1-4.

Ozdas, A., Shiavi, R. G., Silverman, S. E., Silverman, M. K., & Wilkes, D. M. (2004). Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near term suicidal risk. *IEEE Transactions on Biomedical Engineering, 51*(9), 1530-1540.

Park, H., Felty, R., Lormore, K., & Pisoni, D. B. (2010). Presto: Perceptually robust English sentence test: Open-set-design, philosophy, and preliminary findings. *Journal of the Acoustical Society of America, 127*(3), 1958.

Quatieri, T. F., Brady, K., Messing, D., Campbell, J. P., Campbell, W. M., Brandstein, M. S., Weinstien, C. J., Tardelli, J. D., & Gatewood, P. D. (2006). Exploiting nonacoustic sensors for speech encoding. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(2), 533-544.

Quatieri, T. F., Williamson, J. R., Smalt, C. J., Perricone, J., Patel, T., Brattain, L., Helfer, B., Mehta, D., Palmer, J., Heaton, K., Eddy, M., & Moran, J. (2017). Multimodal biomarkers to discriminate cognitive state. In *The Role of Technology in Clinical neuropsychology*, Vol. 409.

Rao, H. M., Smalt, C. J., Rodriquez, A., Wright, H. M., Mehta, D. D., Brattain, L. J., Edwards, H. M., Lammert, A., Heaton, K. J., & Quatieri, T. F. (2020). Predicting cognitive load and operational performance in a simulated marksmanship task. *Frontiers in Human Neuroscience.*

Smalt, C., Rao, H., & Ciccarelli, G. (2021). mit-ll/signal-acquisition-modules-for-lab-streaming-layer: v1.0

Talkar, T., Yuditskaya, S., Williamson, J. R., Lammert, A., Rao, H., Hannon, D., O'Brien, A., Vergara-Diaz, G., DeLaura, R., Sturim, D., Ciccarelli, C., Zafonte, R., Palmer, J., Bonato, P., & Quatieri, T. (2020). Detection of subclinical mild traumatic brain injury (mTBI) through speech and gait. *Proceedings of Interspeech 2020*, 135-139.

Quatieri, T., Messing, D., Brady, K., Campbell, W., Campbell, J., Brandstein, C., Weinstein, J., Tardelli, J., Gatewood, P. (2003). Exploiting nonacoustic sensors for speech enhancement. *Workshop on Multimodal User Authentication. Citeseer*.

Williamson, J. R., Bliss, D. W., Browne, D. W., Indic, P., Bloch-Salisbury, E., & Paydarfar, D. (2011). Using physiological signals to predict apnea in preterm infants. *Conference Record of the 45th Asilomar Conference on Signals, Systems, and Computers, IEEE*, 1089-1102.

Williamson, J. R., Fischl, K., Dumas, A., Hess, A., Hughes, T., & Buller, M. J. (2013). Individualized detection of ambulatory distress in the field using wearable sensors. *IEEE International Conference on Body Sensor Networks*, 1-6.

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Ciccarelli, G., & Mehta, D. D. (2014). Vocal and facial biomarkers of depression based on motor incoordination and timing. *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 65-72.

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Horwitz, R., Yu, B., & Mehta, D. D. (2013). Vocal biomarkers of depression based on motor incoordination. *Proceedings of the 3rd International Workshop on Audio/Visual Emotion Challenge*, 41-48.

Yu, B., Quatieri, T. F., Williamson, J. R., & Mundt, J. C. (2014). Prediction of cognitive performance in an animal fluency task based on rate and articulatory markers. *Fifteenth Annual Conference of the International Speech Communication Association.*

Yu, B., Quatieri, T. F., Williamson, J. R., & Mundt, J. C. (2015). Cognitive impairment prediction in the elderly based on vocal biomarkers. *Sixteenth Annual Conference of the International Speech Communication Association.*

U.S. Army Aeromedical Research Laboratory
Fort Rucker, Alabama

All of USAARL's science and technical information documents are available for download from the Defense Technical Information Center.

https://discover.dtic.mil/results/?q=USAARL

Army Futures Command
U.S. Army Medical Research and Development Command