Audio, Visual, and Electrodermal Arousal Signals as Predictors of Mental Fatigue Following Sustained Cognitive Work

James R. Williamson[1], Kristin J. Heaton[2], Adam Lammert[3], Katherine Finkelstein[4],
Doug Sturim[1], Christopher Smalt[1], Gregory Ciccarelli[1], and Thomas F. Quatieri[1]

Abstract— **Lapses in vigilance and slowed reactions due to mental fatigue can increase risk of accidents and injuries and degrade performance. This paper describes a method for rapid, unobtrusive detection of mental fatigue based on changes in electrodermal arousal (EDA), and changes in neuromotor coordination derived from speaking. Twenty-nine Soldiers completed a 2-hour battery of cognitive tasks intended to induce fatigue. Behavioral markers derived from audio and video during speech were acquired before and after the 2-hour cognitive load tasks, as was EDA. Exposure to cognitive load produced detectable increases in neuromotor variability in speech and facial measures after load and even after a recovery period. A Gaussian mixture model classifier with cross-validation and fusion across speech, video, and EDA produced an accuracy of AUC=0.99 in detecting a change in cognitive fatigue relative to a personalized baseline.**

I. INTRODUCTION

Fatigue can be defined as a psychophysiological state in which capacity to function (physically and/or mentally) is diminished through exertion [1]. Mental fatigue is associated with an increased risk of accidents and injuries due to inattention, impaired decision making, and degraded motor performance [2]. Accurate detection of fatigue is therefore critical before errors are made and accidents occur. At present, there are few methods available that provide rapid, objective and yet unobtrusive evaluation of mental state for use in military operational and training environments [3].

Mental fatigue is typically quantified using subjective ratings of effort, motivation, mood and alertness, objective measures of cognitive and physical performance, and physiological measures. As demands on mental resources increase, self-ratings of tiredness [4] and perceived effort [5] increase, while ratings of motivation to perform [6] and alertness [7] decrease. However, subjective measures are limited by not accounting for individual reporting bias or motivation, leading to possible under or over-estimation of perceived workload [8]. Objective performance measures may require interruption of the task at hand to explicitly elicit verbal or motor responses to discrete stimuli. Speech and facial expression can provide an objective, continuous measure of performance that does not interfere with the primary task because it can be captured unobtrusively during conversation.

Furthermore, there is a neurobiological link between cognitive state and speech production. Mental fatigue develops over time from sustained cognitive demands and is associated with a decrease in cognitive resources available for planning, executive functioning, attention, and fine- and gross-motor activity. Mental workload modifies a finely tuned neural regulatory system interconnecting motor and non-motor areas [9]. Mental fatigue desynchronizes brainwave oscillations that are responsible for functional connectivity across these neural regions [9], resulting in loss of coordination across components of motor activities such as speech production. In addition, mental workload reduces processing resources allocated to motor tasks such as speech production [10], [11]. Speech markers with particular relevance for assessment of mental fatigue include prosody (e.g. pitch variation) [10], [12] and formant structure [11].

Motivated by these observations, we leverage a class of behavioral speech biomarkers that measure speech articulatory coordination (SAC) and that are hypothesized to be sensitive to changes in neuromotor coordination [13], [14]. This work differs from our previous speech/fatigue study [14] with a) a different, larger data collection, b) new audio/video SAC features, c) a new, complementary modality (EDA), and d) using only pre/post speech samples rather than recordings while under mental workload throughout the experiment.

II. METHODS

A. Experimental Protocol

The protocol was approved by the U.S. Army Research Institute of Environmental Medicine Institutional Review Board (USARIEM IRB) and the U.S. Army Medical Research and Materiel Command IRB. The investigators adhered to the policies for protection of human subjects as prescribed in Army Regulation 70-25 and the provisions of 32 CFR Part 219. All participants were briefed on study procedures and gave informed consent prior to the research study. The subjects were 29 active duty Soldiers, predominantly male (83%) with ages 18-26 yrs (mean: 20.21; SD: 2.21).

The 5-day study consisted of an initial screening (Day 1), training (Days 2-4) and experimental procedures (Day 5). In the experiment, speech was collected in three sessions: immediately prior to two-hours of cognitive load (Pre),

1. Massachusetts Institute of Technology, Lincoln Laboratory, Lexington, MA 02421
2. United States Army Research Institute of Environmental Medicine, Natick, MA 01760
3. Worcester Polytechnic Institute, Worcester, MA 01609
4. Seton Hall University, South Orange, NJ 07079

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.

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

immediately following load (PL1), and 15 minutes after the conclusion of PL1 (PL2).

B. Signal Acquisition

Speech was recorded using a standardized protocol including i) read speech (“The Rainbow” [15]), ii) sustained vowels (repeat and hold each of 5-6 vowels for 10 seconds), iii) repeated phonemes (repeat the sounds “pa-ta-ka” as many times as possible in one breath), iv) free speech (respond to open-ended questions such as “Describe a vacation you have taken.”), and v) facial expressions (react to cartoon drawings with contextually appropriate facial and verbal expressions [16]).

Pulse, EDA, and skin temperature data were measured, but only EDA results are reported here because pulse and temperature data were degraded in multiple participants. EDA was available for all but one participant. All signals were acquired using the MindMedia (Herten, Netherlands) NeXus-10 laboratory biofeedback system. Skin conductance was collected using EDA electrodes on the palmar surface of the medial phalange of the first/index finger and third/ring finger of the left hand using a thin Velcro strip.

Audio signals were acquired using a lapel-based (DPA 4061-BM) and a boom microphone (Sennheiser ME66). Video recordings were captured using a high-definition video camera (Canon XA20; Melville, New York). Audio was recorded at 96 kHz, and video was recorded at 30 frames per second (FPS). The read and free speech passages, which are 107 and 89 seconds in duration on average, were selected for analysis because previous research has shown that SAC features are most effective for long duration speech. The EDA signals were analyzed only during these speech passages so that the signals were compared under similar conditions.

C. Low-level features

Raw audio was first transformed into low-level features. Different acoustic frequencies are emphasized or attenuated in a time-varying way as speech articulators move. These frequency patterns were captured using formant frequencies and mel-frequency cepstral coefficients. The three lowest formant frequencies were tracked and extracted every 10 ms from the audio signal using the KARMA software tool [17]. Delta-formants (dFormants), the discrete-time derivatives of the formants, were also computed. A total of 16 delta-MFCCs (dMFCCs), which are the discrete time derivatives of Mel-frequency cepstral coefficients (MFCCs), were extracted using openSMILE [18]. Visual speech-related movements from video of the face during speaking were also captured. Facial action unit (FAU) features were extracted using the software tool openFace [19]. The intensity of 17 FAUs from each video frame were estimated along with delta-FAUs (dFAUs), which are their discrete-time derivatives across video frames.

EDA signals were processed using the software tool cvxEDA in the NeuroKit python toolkit [20], which generates the raw EDA signal and its phasic and tonic components. In this approach phasic activity is assumed to be superimposed on a slowly varying tonic activity that has a spectrum below 0.05 Hz.

D. High-level features

It is hypothesized that changes in neurological functioning, including temporary decrements due to mental fatigue, alter the neuromotor timing and coordination of speech articulation. Measures have been developed that quantify the level of motor coordination through correlation patterns among different channels of each low-level multichannel feature set [13]. Features that capture levels of coordination from speech, referred to as speech articulatory coordination (SAC) features, were next extracted from the low-level features.

SAC features are the eigenspectra of channel-delay correlation matrices that are computed from low-level multichannel signals derived from speech. The correlation matrices are constructed using time-delay embedding at multiple delay scales. Specifically, a channel-delay correlation matrix at delay scale j is computed as

$$\mathbf{R_j} = \begin{bmatrix} R_{1,1}(j) & \dots & R_{1,M}(j) \\ \vdots & \ddots & \vdots \\ R_{M,1}(j) & \dots & R_{M,M}(j) \end{bmatrix} \tag{1}$$

where M is the number of low-level feature channels. Each submatrix $R_{c_1,c_2}(j)$ contains the set of correlations between channels c_1 and c_2 at scale j,

$$R_{c_1,c_2}(j) = \begin{bmatrix} r_{1,1}(j) & \dots & r_{1,N}(j) \\ \vdots & \ddots & \vdots \\ r_{N,1}(j) & \dots & r_{N,N}(j) \end{bmatrix}_{c_1,c_2} \tag{2}$$

N is the number of delays per channel and $[r_{d_1,d_2}(j)]_{c_1,c_2}$ is the correlation, at scale j, between channel c_1 at delay d_1 with channel c_2 at delay d_2.

Complexity within the correlation matrix is quantified using the matrix eigenspectrum. A greater concentration of weight in the largest eigenvalues indicates lower complexity, whereas a greater concentration of weight in the smaller eigenvalues indicates higher complexity. Fig. 1 diagrams how the SAC measures were computed from speech formant frequencies.

Notional effect sizes of the eigenspectra features, which are in descending order of size, are plotted in Fig. 1. The pattern of effect sizes as a function of eigenvalue index indicates how different types of neuromotor degradation can be manifested in speech, either through more erratic (red line) or more simple (black line) speech articulation. Previous work has shown that certain sources of degraded mental state, such as major depressive disorder, produce coordination patterns that are less complex [21]. Other sources of degraded mental state produce patterns that appear more complex, which in the context of degraded mental state can be interpreted as more erratic. These sources include cognitive load [13], physical fatigue combined with exposure to heat or altitude [22] and mental fatigue [14]. We speculate that the observed correlations in deviations of the eigenspectrum with a variety

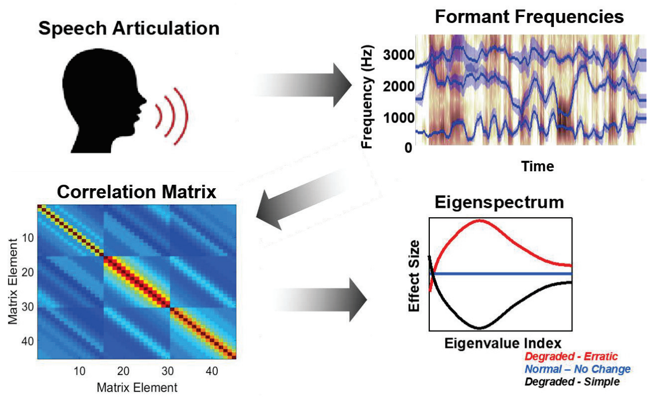

Fig. 1. Computation of SAC features from speech formants.

of neuropsychological changes are driven by changes in the underlying neuroregulatory systems, as discussed in the Introduction.

In the present analysis, the effects of cognitive fatigue on speech neuromotor coordination are evaluated using SAC measures computed from five feature sets: formants, dFormants, dMFCCs, FAUs and dFAUs. Time delay embedding is done at four delay scales, with delay spacings of 1, 3, 7 and 15 frames. These correspond to time delays of 10, 30, 70 and 150 ms per delay for the 100 Hz audio-based features and time delays of 33, 100, 230, and 500 ms per delay for the 30 Hz video-based features. There were 15 time delays per scale, ranging from a delay of zero up to a delay of 14 times the delay-scale spacing. The number of channels for the low-level features are: $M = 3$ (formants, dFormants), $M = 16$ (dMFCC), and $M = 17$ (FAUs, dFAUs). The EDA high level features are the mean and standard deviation of the raw, phasic, and tonic EDA signals.

E. Machine Learning

The primary goal of this work is to detect cognitive fatigue, assumed to exist post two hour mental workload but not before, based on SAC and EDA features. To test this capability, the dataset of 29 participants was randomly partitioned into five cross-validation test folds of five or six participants each. For each of those folds, the classifier was trained on the remaining subjects. The SAC features consist of eigenvalue vectors concatenated across four delay scales. The EDA features consist of six statistical features.

The SAC and EDA measures were normalized to have zero mean within subject prior to dimensionality reduction via principal component analysis (PCA). This was done because the detection problem is to find within-subject changes rather than across-subject variation. The substantial, natural, person-to-person variability in absolute feature values further motivates this normalization procedure. Additionally, we believe this step is perfectly in line with a fieldable algorithm because speech is so easily collected that the idea of voice banking a person's natural speech under different conditions may become as accepted a practice as charting blood pressure, heart rate, and weight.

Specifically, prior to PCA, the mean of each feature element was computed across the three sessions, per subject, and subtracted. Then, the training set features were z-scored, across all the subjects, and PCA was applied. This within-subject feature subtraction was also applied to the held-out test subjects. Then, the z-scoring and PCA transforms from the training set were applied to the held-out test data.

The single free parameter for each feature set is the number of principal components, K, that are extracted. K is selected using another level of cross-validation within the training set in order to produce unbiased detection accuracy estimates. Within each training fold, a second level of four-fold cross-validation is done, with potential values of K varied between one and 15, and the value of K is selected that produces the highest average area under the ROC curve (AUC) within the training fold. In addition, if AUC < 0.6 then this feature set is not included in any fusion across feature sets or across speech tasks on the associated test fold.

The classification method first creates a probability density model for each feature set from the training data using unsupervised learning that is inclusive of all three experimental sessions (Pre, PL1, PL2). Next, two-class conditional probability density models are created by shifting the probability density toward the features from the Pre (Class 1) and PL1 (Class 2) sessions. Given a test datum, the log-likelihood ratio of the two models is the output score. The classification method was implemented by fitting a Gaussian mixture model (GMM) to the training data and adapting this unsupervised model to the two output classes using the procedure described in [23]. The GMM uses 10 Gaussian components with diagonal covariance matrices, which are fit to the training data of both output classes using five iterations of the Expectation Maximization algorithm. In addition, 10 independent GMMs are trained (with different random initializations), and the GMM likelihoods for each class are summed over the ensemble of models before computing the final output score from the log-likelihood ratio of the sums.

III. Results

Fig. 2 shows effect sizes associated with mental workload observed for the Formant (top left), dFormant (top right), FAU (bottom left), and dFAU (bottom right) SAC features. These plots show the effect size of features computed from read speech and free speech in the PL1 session compared to those from the Pre session, all at the first delay scale. The effect size patterns are broadly similar to the notional pattern in Fig. 1 (lower right) that is indicative of more erratic speech.

Accuracy in detecting mental fatigue, based on feature changes such as those shown in Fig. 2, is quantified using the AUC, which is computed from a union of prediction scores across the five cross-validation test folds. Table I shows the AUC results for each speech SAC feature set on both read and free speech, and testing on the PL1 and PL2 sessions. dFormant and dFAU features are effective on both read and free speech. dMFCC and FAU features are effective on read speech only, whereas formant features are

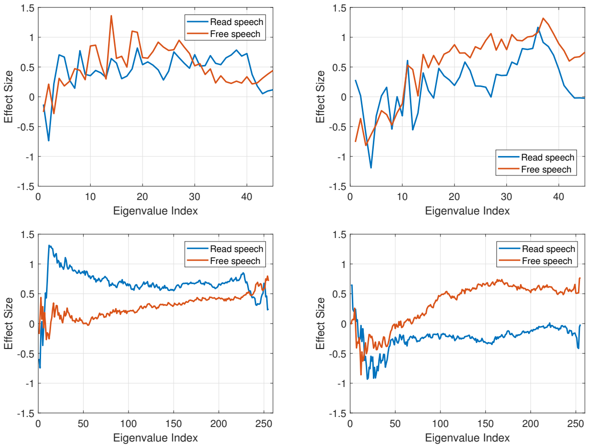

Fig. 2. Effect sizes of Formant (top left), dFormant (top right), FAU (bottom left), and dFAU (bottom right) SAC features for read and free speech at the first delay scale.

TABLE I

ACCURACY IN DETECTING COGNITIVE FATIGUE FROM INDIVIDUAL FEATURE SETS.

Feature	Read AUC		Free AUC	
	PL1	PL2	PL1	PL2
Formant	0.53	0.56	0.77	0.71
dFormant	0.79	0.75	0.72	0.70
dMFCC	0.73	0.73	0.53	0.49
FAU	0.71	0.61	0.47	0.59
dFAU	0.70	0.57	0.75	0.71

TABLE II

ACCURACY IN DETECTING COGNITIVE FATIGUE BY FUSING ACROSS FEATURE SETS AND SPEECH PASSAGES.

Modality	Read AUC		Free AUC		R, F AUC	
	PL1	PL2	PL1	PL2	PL1	PL2
Audio	0.80	0.77	0.73	0.66	0.85	0.80
Video	0.78	0.60	0.75	0.73	0.84	0.74
A,V	0.85	0.77	0.77	0.74	0.93	0.86
EDA	0.91	0.88	0.82	0.84	0.91	0.91
A,V, EDA	0.96	0.92	0.91	0.90	0.99	0.97

effective on free speech only. Overall, the AUC values show that individual SAC feature sets are moderately effective at detecting fatigue, and that detection performance declines somewhat in PL2.

Table II shows how detection performance improves with fused combinations of SAC predictions. These fusions are shown both across different feature sets, and across the read and free speech passages. Audio and Video SAC features achieve similar performance, with AUCs of 0.85 and 0.84 by combining across read and free speech. Fusing across audio and video features, and across read and free speech produces AUC of 0.93 for PL1. SAC features show a moderate decline in accuracy between PL1 and PL2. Finally, fusing across all feature modalities produces extremely high detection accuracy, with AUC=0.99.

IV. DISCUSSION

We compared the efficacy of speech, facial expression, and EDA, individually and when fused, at detecting a change in fatigue relative to a personalized baseline that is produced by two hours of cognitive work. The detection was done both immediately following the work, and after a short rest period.

In the two post-load sessions, speech and facial features showed an increase in the independence and variability of underlying motor components of each modality, consistent with previous findings in smaller datasets on cognitive load [13] and cognitive fatigue [14]. An unexpected finding with both modalities was the small difference in accuracy in detecting the two post-load conditions, indicating a strong lingering effect of the two hours of mental activity.

Audio and video modalities produced similar accuracy in detecting fatigue on both read and free speech. For both modalities, better detection performance was found on read speech than free speech, with the best performance found by fusing across both speech types. Finally, the highest detection accuracy of AUC=0.99 was found by fusing across the audio, video, and EDA modalities, as well as across the two types of speech. These results indicate that speech articulatory coordination, estimated from audio and/or video, has strong potential for use as a practical indicator of cognitive fatigue. Furthermore, though neuromotor coordination analysis was applied to speech motor control, the concept may generalize to other signals derived from neural function.

REFERENCES

[1] R. O. Phillips, “A review of definitions of fatigue–and a step towards a whole definition,” *Transportation research part F: traffic psychology and behaviour*, vol. 29, pp. 48–56, 2015.

[2] A. Williamson, D. A. Lombardi, S. Folkard, J. Stutts, T. K. Courtney, and J. L. Connor, “The link between fatigue and safety,” *Accident Analysis & Prevention*, vol. 43, no. 2, pp. 498–515, 2011.

[3] S. P. Proctor, K. J. Heaton, H. R. Lieberman, C. D. Smith, E. N. Edens, A. Kelley, T. J. Balkin, V. Capaldi, T. J. Doty, and P. J. Quartana, “Military cognitive performance and readiness assessment initiative,” Tech. Rep., ARMY RESEARCH INST OF ENVIRONMENTAL MEDICINE, Natick, MA USA, 2017.

[4] M. A. Boksem and M. Tops, “Mental fatigue: costs and benefits,” *Brain research reviews*, vol. 59, no. 1, pp. 125–139, 2008.

[5] B. Pageaux and R. Lepers, “Fatigue induced by physical and mental exertion increases perception of effort and impairs subsequent endurance performance,” *Frontiers in physiology*, vol. 7, pp. 587, 2016.

[6] M. A. Boksem, T. F. Meijman, and M. M. Lorist, “Mental fatigue, motivation and action monitoring,” *Biological psychology*, vol. 72, no. 2, pp. 123–132, 2006.

[7] D. van der Linden, S. A. Massar, A. F. Schellekens, B. A. Ellenbroek, and R.-J. Verkes, “Disrupted sensorimotor gating due to mental fatigue: preliminary evidence,” *International journal of psychophysiology*, vol. 62, no. 1, pp. 168–174, 2006.

[8] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. Van Gerven, “Cognitive load measurement as a means to advance cognitive load theory,” *Educational psychologist*, vol. 38, no. 1, pp. 63–71, 2003.

[9] A. Ishii, M. Tanaka, and Y. Watanabe, “Neural mechanisms of mental fatigue,” *Reviews in the Neurosciences*, vol. 25, no. 4, pp. 469–479, 2014.

[10] J. D. Harnsberger, R. Wright, and D. B. Pisoni, “A new method for eliciting three speaking styles in the laboratory,” *Speech communication*, vol. 50, no. 4, pp. 323–336, 2008.

[11] S. E. Lively, D. B. Pisoni, W. Van Summers, and R. H. Bernacki, "Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2962–2973, 1993.

[12] K. Huttunen, H. Keränen, E. Väyrynen, R. Pääkkönen, and T. Leino, "Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights," *Applied ergonomics*, vol. 42, no. 2, pp. 348–357, 2011.

[13] T. F. Quatieri, J. R. Williamson, C. J. Smalt, J. Perricone, T. Patel, L. Brattain, B. Helfer, D. Mehta, J. Palmer, K. Heaton, et al., "Multimodal biomarkers to discriminate cognitive state," in *The Role of Technology in Clinical Neuropsychology*. Oxford University Press, Oxford, 2017.

[14] J. Sloboda, A. Lammert, J. Williamson, C. Smalt, D. D. Mehta, C. I. Curry, K. Heaton, J. Palmer, and T. Quatieri, "Vocal biomarkers for cognitive performance estimation in a working memory task," in *Proc. Interspeech 2018*, 2018, pp. 1756–1760.

[15] G. Fairbanks, "The rainbow passage," *Voice and articulation drillbook*, vol. 2, 1960.

[16] B. Kolb, B. Wilson, and L. Taylor, "Developmental changes in the recognition and comprehension of facial expression: Implications for frontal lobe function," *Brain and Cognition*, vol. 20, no. 1, pp. 74–84, 1992.

[17] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1732–1746, 2012.

[18] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[19] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," Tech. Rep., CMU-CS-16-118, CMU School of Computer Science, 2016.

[20] A. Greco, G. Valenza, and E. P. Scilingo, *Advances in Electrodermal Activity Processing with Applications for Mental Health*, Springer, 2016.

[21] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 65–72.

[22] J. R. Williamson, T. F. Quatieri, A. C. Lammert, K. Mitchell, K. Finkelstein, N. Ekon, C. Dillon, R. Kenefick, and K. Heaton, "The effect of exposure to high altitude and heat on speech articulatory coordination.," in *Interspeech*, 2018, pp. 297–301.

[23] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.