Interspeech 2018
2-6 September 2018, Hyderabad

Vocal biomarkers for cognitive performance estimation in a working memory task*

Jennifer Sloboda[1], *Adam Lammert*[1], *James Williamson*[1], *Christopher Smalt*[1], *Daryush D. Mehta*[1,2], *COL Ian Curry*[3], *Kristin Heaton*[4], *Jeffrey Palmer*[1], *Thomas Quatieri*[1]

[1]MIT Lincoln Laboratory, Lexington, MA, USA
[2]Massachusetts General Hospital, Center for Laryngeal Surgery and Voice Rehabilitation
[3]US Army Aeromedical Research Laboratory
[4]US Army Research Institute of Environmental Medicine
Jennifer.Sloboda@ll.mit.edu

Abstract

The ability to non-invasively estimate cognitive fatigue and workload as contributing factors to cognitive performance has value for planning and decision making surrounding human participation in cognitively demanding situations and environments. Growing evidence supports the use of speech as an effective modality for assessing cognitive fatigue and workload, while also being operationally appropriate in a wide variety of environments. To assess ability to discriminate changes in cognitive fatigue and load from speech, features that measure speech onset time, speaking rate, voice quality, and vocal tract coordination from the delta-Mel-cepstrum are evaluated on two independent data sets that employ the same auditory working memory task. Feature effect sizes due to fatigue were generally larger than those due to load. Speech onset time, speaking rate, and vocal tract coordination features show strong potential for speech-based fatigue estimation.

Index Terms: vocal biomarkers, cognitive fatigue, cognitive load, phoneme and pause duration, articulatory coordination

1. Introduction

Cognitive performance is affected both by a person's cognitive workload and their cognitive fatigue. Cognitive workload refers to the mental demand experienced for a task, determined by a person's effort level and cognitive capacity, intrinsic task difficulty, and presence of extraneous distractors [1]-[3]. Cognitive fatigue is a subjective experience of mental weakness, closely associated with increased somnolence [2], but also reduction in cognitive capacity, which is a slowing or lessening of ability to perform cognitive tasks [4], including reduction in working memory capacity, reduced reaction time and insensitivity to external stimuli.

The ability to estimate cognitive fatigue and workload independently, or as joint contributing factors to cognitive performance, has value for planning and decision making surrounding human participation in cognitively demanding situations and environments. Moreover, the ability to estimate these factors noninvasively is critical, because such situations often cannot tolerate interruptions or distractions. For instance, the pressures of military training and operations demand high cognitive workload, and high levels of cognitive fatigue are likely. Accurate estimates of cognitive status may inform mission planning and command decisions, as well as the role for and proper interaction with augmenting/autonomous systems. At the same time, well-established tests, such as the Psychomotor Vigilance Task (PVT), are inappropriate for such situations because they are obtrusive, requiring individuals to disengage from their primary task and attend to the test itself. Growing evidence supports the use of speech as an effective modality for assessing cognitive fatigue and workload, while also being operationally appropriate in a wide variety of environments [5]-[10].

Specifically, prior study of intra-individual change in speech over a period of sustained wakefulness found increased total speech time, mean pause length, and total signal time in fatigued speech [5]. Prior study of high and low cognitive load discrimination showed greater than 90% accuracy using features capturing articulatory source coordination [6], [11]. Building off this work, we sought to investigate the differentiability of both high and low cognitive fatigue and load conditions from vocal biomarkers capturing timing, voice quality and articulatory coordination.

The paper is organized as follows. Section 2 describes the data collection from two independent studies of speech during an auditory working memory task, including derivation of cognitive load and fatigue level labels. Section 3 describes speech feature extraction and effect size methodologies. Section 4 presents the effects of cognitive load and fatigue on feature distributions in each study, compares the results, and discusses insights gained. Section 5 provides conclusions and directions for future work

2. Methods

This paper analyzes data from two independent experiments, which employ the same auditory working memory task, to test cognitive fatigue and load discrimination from audio-based speech features. Study 1 is an approximately two-hour exercise proctored in laboratory conditions. Study 2 is a daily "take-

Table 1: Key study differences

Study	Study 1	Study 2
Conditions	Laboratory	non-laboratory
Collect frequency	one-time	daily
# consecutive trials	324 (total)	10 (per day)
Duration	~2 hours	5-23 days**
High/low load	n / n-2 digits*	5 / 0 digits
High/low fatigue	second / first half of trials	Low / high reaction speed

*where n is a subject-specific calibrated value **see Table 3

*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. FA8721-05-C-0002 and/or 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.

10.21437/Interspeech.2018-2418

home" exercise repeated for several days. Both experimental protocols test auditory working memory by requiring the subject to recall a sentence while holding a number of digits in memory. This task is henceforth referred to as the digit span task [12]. A single trial of the digit span task comprises a subject hearing a string of digits, then a sentence, then two tones eliciting spoken recall of the sentence, and then the digits. The multi-talker PRESTO sentence database is used for sentence stimuli [13]. Speech features are extracted from the repeat sentence interval.

Different load conditions are induced by changing the number of digits being held in working memory. The Study 1 protocol induces cognitive fatigue by consecutive repetition of 324 digit span trials, resulting in increasing stress on auditory working memory (thus increasing cognitive fatigue) over time. Study 2 records subjects' "naturally occurring" cognitive fatigue level via a psychomotor vigilance test (PVT) taken after the working memory task. Following are details about each experiment protocol.

2.1. Study 1 protocol

Data was collected from 16 native English speaking subjects (8 male, 8 female), with mean age 36.3 years, in laboratory conditions. A working memory-based protocol approved by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES) was followed. After setup and training, each subject engaged in 108 digit span task trials at each of three cognitive load difficulty levels [14], [15], [12]. Each difficulty level used the same set of 108 unique sentences and the order of the resulting 324 trials was randomized. An initial calibration test was done to assess each subjects' ability in the working memory test. The maximum number of digits a subject is able to recall, with 71% accuracy, was estimated using a two-up one-down adaptive tracking algorithm [14]. This maximum number, n, was determined to be: 4 (for four subjects), 5 (six subjects), 6 (four subjects), and 7 (two subjects). For the analyses presented in this paper, the trials from two load difficulty levels (n and n-2) were used, 216 trials in total. Audio data was collected with a DPA acoustic lapel microphone (with a Roland Octa-Capture audio interface) and a 44 kHz sampling rate.

2.2. Study 2 protocol

Data was collected on a daily basis from four male, native English speaking subjects, with mean age 35.5 years, in the location of their choosing (often a home or work environment). Duration and consistency of daily participation varied between subjects (see Table 3). Following an MIT COUHES-approved protocol, subjects performed several tasks, including working memory and reaction time tasks. On an iPod touch (with an audio sampling rate of 44 kHz), subjects completed ten consecutive trials of the digit span task each day: five trials with five digits for recall, and five trials with zero digits for recall (i.e., hear sentence, repeat sentence). A 15-second recording window was used to capture the subjects' repetition of each sentence. Subjects were instructed to take a PVT test, using a custom-built PVT device, soon after completing the digit span exercise. One day of data is referred to as one session.

2.3. Data labeling

Every digit span task trial is labeled as "high" or "low" for both cognitive load and cognitive fatigue conditions. In Study 1, the first half of the trials are labeled as low cognitive fatigue, and the second half as high fatigue. High or low cognitive load conditions were defined as an n or n-2 digit task, respectively, where n is the maximum number of digits which can accurately be recalled by a subject. In Study 2, cognitive fatigue labels are assigned by PVT reaction speeds. A measure of mean reciprocal reaction time is calculated from the PVT data recorded in a session by discarding reaction times outside of 100-500 ms (anticipations or lapses) and taking the average of reciprocal reaction times. The resulting reaction speed values were normalized to zero mean within subject. A k-means clustering algorithm (where $k = 2$) was applied, separating the data into high and low reaction speeds, corresponding to low and high cognitive fatigue level labels for each session. High and low cognitive load conditions are defined as a five digit or zero digit task, respectively.

3. Feature Extraction

3.1. Low-level feature extraction

Preprocessing: In Study 2, due to the fact that audio recordings were done using a non-laboratory protocol, some low quality recordings needed to be filtered out. Trials were discarded if they contained serious recording problems (e.g., a second voice in recording) or incomplete recordings. This procedure resulted in eliminated sessions or fewer than 10 trials in a session. Audio files were denoised using an MMSE-based speech enhancement routine, and denoised files were trimmed to eliminate pre- and post-speech silence using a custom speech activity detector.

Phonemes: An automatic phoneme recognition algorithm [16], for Study 1, and a forced alignment algorithm, for Study 2, were used to detect phonetic boundaries in each sentence recording. Each segment was labeled with one of 41 phoneme classes: 40 ARPABET phonemes, plus a custom class for first and last pause.

Cepstral peak prominence (CPP): CPP is an acoustic measure with strong reported correlations to overall dysphonia perception, breathiness, and vocal fold kinematics, which does not rely on an accurate estimate of fundamental frequency [17]. CPP measures the decibel difference between the noise floor and the magnitude of the highest peak in the power cepstrum for quefrencies greater than 2 ms (corresponding to a range minimally affected by vocal tract–related information) [17]. In Study 1, CPP was calculated with a window size of 20 ms, and 40.96 ms in Study 2.

Creaky voice probability (Creak): A creaky voice quality (vocal fry, irregular pitch periods, etc.), is characterized using acoustic measures of low-frequency/damped glottal pulses [18]. A frame-based measure of creak posterior probability is calculated as in [6], using a fusion of short-term power (4 ms windowing), intraframe periodicity (32 ms window), inter-pulse similarity, and two measures of the degree of sub-harmonic energy (reflecting the presence of secondary glottal pulses) and the temporal peakiness of glottal pulses with long period.

Harmonics-to-noise ratio (HNR): Applying the techniques described in [17], a spectral measure of short-time harmonics-to-noise ratio, which is the ratio (in dB) of the power of the decomposed harmonic signal and the power of the decomposed speech noise signal, was computed. In Study 1, a measure of

HNR is computed with a window size of 20 ms. In Study 2, two different measures of HNR are computed (and included in the aggregate vocal features): one with a window size four times the period and one with a window size of 40 ms.

Low-to-high frequency spectral ratio (LH ratio): LH ratio is calculated as the difference between spectral power below and above a cutoff frequency, for 8 cutoff frequencies: every 1kHz from 1-8kHz [19]. A 40 ms window size was used.

Delta Mel-frequency cepstral coefficients (dMFCCs): 16 MFCC channels are computed and delta MFCCs [20], generated by differencing across consecutive 10 ms frames, are used to characterize velocities of vocal tract spectral magnitudes.

3.2. High-level features

Speech onset time: Using the phoneme boundaries, the duration of the first pause before speech onset is recorded as speech onset time. For Study 2, the mean speech onset time per session is calculated.

Overall speaking rate: Overall speaking rate is calculated from the phoneme durations as total number of phonemes divided by sentence duration, excluding first pause.

Aggregate vocal features: Low-level vocal features (CPP, creak, HNR, and LH ratio) were calculated on a frame-by-frame basis for each audio recording. Feature values – from one trial, in Study 1, and from all sentences in each load condition a session, in Study 2 - are pooled and summary statistics (mean, median, standard deviation, skewness, and kurtosis) are calculated. PCA was employed to reduce dimensionality of the features to one.

dMFCC correlation structure: Measures of the structure of correlations among low-level speech features have been applied in the estimation of depression, performance associated with dementia, and changes in cognitive performance associated with mild traumatic brain injury [6]. Correlation structure refers to the rank-ordered eigenvalues of correlation matrices and, thus, shape of distribution independent of data axes. The differences in dMFCC eigenspectra patterns due to high or low cognitive fatigue and/or load reflects the effect of each condition on coordination of vocal tract trajectories. PCA was employed to reduce dimensionality of the features to one.

3.3. Effect size calculation

Effect sizes of high-level features are used to quantify the difference in feature distributions across fatigue and load conditions. Effect size quantifies the difference in the means of two distributions relative to their standard deviations [21]. As such, it is advantageous for analyzing modest datasets (like Study 2) and observations of inter-subject variability.

Hedges' g effect size across classes A and B is given as:

$$g = \frac{\overline{x_A} - \overline{x_B}}{SD^*_{pooled}} \quad (1)$$

where the pooled standard deviation (for two independent samples) is defined as [22]:

$$SD^*_{pooled} = \sqrt{\frac{(n_A-1)\sigma_A^2+(n_B-1)\sigma_B^2}{n_A+n_B-2}} \quad (2)$$

For normally distributed populations, an effect size equal to one may be interpreted as a difference in group means equal to one standard deviation. The product of effect size and size of the studied population may be interpreted as a test of significance.

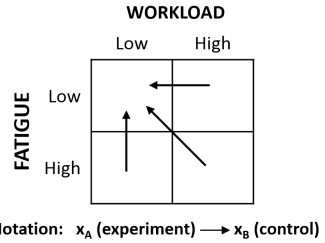

Figure 1: Diagram of Cognitive Conditions.

Effect size of each combination of high condition relative to the low load low fatigue condition was calculated: high load-low fatigue, low load-high fatigue, and high load-high fatigue.

4. Results

Results from Study 1 are presented in aggregate, for the 16 subjects in the study. In Study 1, there were 216 valid trials from each subject, resulting in 54 trials per condition. Table 2 summarizes the aggregate effect sizes for the four different speech features applied to the three experimental conditions, which are summarized in Figure 1. Load changes alone do not produce feature changes with significant effect sizes. For three of the features, fatigue and fatigue combined with load do produce significant changes (shown in bold font). These changes appear to be due to fatigue alone, based on the similarity of effect sizes for fatigue alone and for fatigue combined with load. The changes are an increase in the speech onset time, a decrease in the speaking rate, and an increase in the DMFCC first principal component (which corresponds to larger values in the low-rank eigenvalues).

Table 2: Study 1 effect sizes

Feature	Condition	Effect Size	P-value
Speech Onset Time	Load	-0.10	0.04
	Fatigue	**0.34**	**0.00**
	Load & Fat.	**0.30**	**0.00**
Speaking Rate	Load	-0.08	0.09
	Fatigue	**-0.20**	**0.00**
	Load & Fat.	**-0.20**	**0.00**
Voice Features	Load	0.04	0.43
	Fatigue	-0.03	0.51
	Load & Fat.	0.03	0.51
DMFCC Corr. Struct.	Load	0.01	0.82
	Fatigue	**0.28**	**0.00**
	Load & Fat.	**0.28**	**0.00**

These changes are consistent with psychomotor retardation. For example, in previous work on the effects of depression on speech features, similar relations were found between depression severity and DMFCC eigenvalues and phoneme-based speaking rate [6], [17]. These changes have been hypothesized to be due to psychomotor retardation, which is a prominent symptom of major depressive disorder. To our knowledge, speech onset time has not previously been utilized as an indicator of psychomotor retardation.

In Figure 2, the distribution of effect sizes in the 16 subjects are plotted for the four speech features, with the aggregate effect size shown with a red bar. Notice that the effect size for many individual subjects has the opposite sign as the aggregate effect size. These distributions provide the appropriate context for studying the effect sizes found on an individual basis in Study 2.

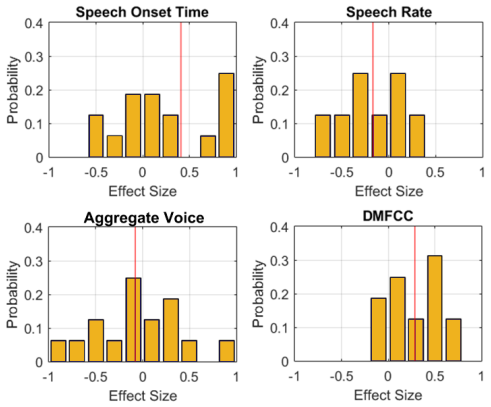

Figure 2: Effect size distributions for High Fatigue-High Load condition for Study 1.

The results from Study 2 are presented on a per-subject basis, as a case study on individual variability. Subject 373 is not presented, as there were too few data points from which to calculate a meaningful effect size. In Study 2, the number of valid sessions per subject in each fatigue/load group are shown in Table 3 and the PVT reaction speeds (average reciprocal reaction time) are shown in Table 4 in both raw and normalized (zero-mean) forms. Note that large reaction speeds correspond to small reaction times and lower fatigue conditions.

Table 3: Study 2 valid sessions

Subject ID	**135**	**283**	**315**	**373**
Valid Sessions	22	15	23	5
low load, low fatigue	11	8	10	3
low load, high fatigue	11	7	23	2
high load, low fatigue	11	8	10	3
high load, high fatigue	11	7	13	2

Table 4: *Study 2 PVT reaction speed ranges (s^{-1})*

Subject	**All**	**135**	**283**	**315**	**373**
min (raw)	3.7	4.6	3.7	4.9	4.0
max (raw)	5.7	5.2	4.4	5.7	4.5
min (-mean)	-0.38	-.38	-0.36	-0.33	-0.30
max (-mean)	0.43	0.26	0.31	0.43	0.18

Figure 3 illustrates representative results from Study 2 on an aggregate and individual bases. The aggregate results from Study 1 are also provided for comparison. Study 2 differs from Study 1 in two important respects. First, a smaller number of sessions are used to compute individual and aggregate effect sizes. Second, in Study 1 the fatigue label is based the time period within the experiment in which the trial occurred, whereas in Study 2 the fatigue label is based on the PVT score. Speech onset time appears to be the most consistent feature across the two studies.

Both studies show a range of individual effects, indicating that effective methods for cognitive performance estimation may require individualization, as similar levels of induced

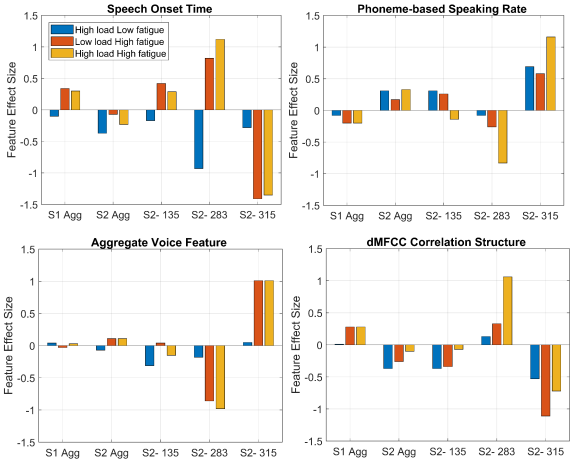

Figure 3: (top left) speech onset time effect size in Study 2, (top right) phoneme-based speaking rate effect size in Study 2, (bottom left) aggregate voice feature effect size in Study 2, and (bottom right) dMFCC correlation structure effect size in Study 2.

fatigue may affect speech patterns differently in different individuals.

5. Conclusions

This paper investigates independent and joint effects of cognitive load and fatigue conditions in voice features capturing timing, voice quality, and articulatory coordination. We find the fatigue effect is typically much larger than the load effect in all features, which has implications for the joint estimation task. More research is necessary to characterize a cognitive performance estimation technique taking individual variability into account.

6. Acknowledgements

Support for this work provided by US Army. The authors would like to thank the following staff at MIT Lincoln Laboratory: Tejash Patel, Darrell Ricke, and Kate Byrd for contributions.

7. References

[1] 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.

[2] P. L. Ackerman, Cognitive fatigue: Multidisciplinary perspectives on current research and future applications, Washington DC, US:APA Press, 2010.

[3] J. Sweller, "Element interactivity and intrinsic, extraneous, and germane cognitive load," *Educational psychology review*, vol. 22, no. 2, pp. 123–138, 2010.

[4] T. B. Sheridan, H. G. Stassen, "Definitions models and measures of human workload" in Mental Workload - Its Theory and Measurement, New York:Plenum Press, 1979.

[5] A. P. Vogel, J. Fletcher, and P. Maruff, "Acoustic analysis of the effects of sustained wakefulness on speech," *Journal of the Acoustical Society of America*, vol. 128, pp. 3747–3756, 2010.

[6] T. Quatieri, J. Williamson, C. Smalt, T. Patel, J. Perricone, D. Mehta et al., "Vocal biomarkers to discriminate cognitive load in

a working memory task"; INTERSPEECH 2015: 15th Annual Conf. of the Int. Speech Communication Assoc., 6-10 September 2015.
[7] B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah, "Speech-based cognitive load monitoring system," Proc. ICASSP, 2008.
[8] B. Yin and F. Chen, "Towards automatic cognitive load measurement from speech analysis," in Human-Computer Interaction. Interaction Design and Usability. Springer Berlin Heidelberg, 2007, pp. 1011–1020.
[9] M. A. Khawaja, N. Ruiz, and F. Cheng, "Think before you talk: An empirical study of relationship between speech pauses and cognitive load," Proc. OZCHI, December 8–12, 2008.
[10] Le, P., J. Epps, Choi, H.C., and Ambikairajah, E. (2010). A study of voice source- and vocal tract-based features in cognitive load classification. *Proceedings of International Conference on Pattern Recognition*, 4516-4519.
[11] 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," *The Role of Technology in Clinical Neuropsychology*, p. 409, 2017.
[12] 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.
[13] Gilbert, Jaimie L., Terrin N. Tamati, and David B. Pisoni. "Development, reliability, and validity of PRESTO: A new high-variability sentence recognition test." *Journal of the American Academy of Audiology* 24.1 (2013): 26-36.
[14] H. Levitt, "Transformed up-down methods in psychoacoustics," *The Journal of the Acoustical society of America*, vol. 49, no. 2B, pp. 467–477, 1971.
[15] Y. Li, Y. Liu, J. Li, W. Qin, K. Li, C. Yu, and T. Jiang, "Brain anatomical network and intelligence," *PLoS Comput Biol*, vol. 5, no. 5, p. e1000395, 2009.
[16] W. Shen, C. White, T. J. Hazen, "A comparison of query-by-example methods for spoken term detection," in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, 2010.
[17] James R. Williamson, Thomas F. Quatieri, Brian S. Helfer, Gregory Ciccarelli, and Daryush D. Mehta. 2014. Vocal and Facial Biomarkers of Depression based on Motor Incoordination and Timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* (AVEC '14). ACM, New York, NY, USA, 65-72.
[18] B. R. Gerratt and J. Kreiman, "Toward a taxonomy of nonmodal phonation," *Journal of Phonetics*, vol. 29, no. 4, pp. 365-381, 2001.
[19] S. N. Awan, N. Roy, M. E. Jetté, G. S. Meltzner, and R. E. Hillman, "Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: Comparisons with auditory-perceptual judgements from the CAPE-V," *Clin. Linguist. Phon.*, vol. 24, no. 9, pp. 742--758, 2010.
[20] T.F. Quatieri, Discrete-time speech signal processing: Principles and Practice, Pearson, 2002.
[21] Cohen J. Hillsdale, NJ: Lawrence Erlbaum; 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd ed
[22] Hedges, L.V. (1981), "Distribution theory for Glass's estimator of effect size and related estimators," *Journal of Educational Statistics*, 6(2): 106–128.
[23] Ellis PD. The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results. Cambridge, UK: Cambridge University Press; 2010.