Classification of Cognitive Load from Speech using an i-vector Framework

Maarten Van Segbroeck[1], *Ruchir Travadi*[1], *Colin Vaz*[1], *Jangwon Kim*[1],
Matthew P. Black[2,3], *Alexandros Potamianos*[1,3], *Shrikanth S. Narayanan*[1,3]

[1]Signal Analysis & Interpretation Laboratory, Univ. of Southern California, Los Angeles, CA
[2]Information Sciences Institute, Univ. of Southern California, Marina del Rey, CA
[3]Behavioral Informatix, LLC, Los Angeles, CA

[1]http://sail.usc.edu, [2]www.isi.edu, [3]www.behavioralinformatix.com

Abstract

The goal in this work is to automatically classify speakers' level of cognitive load (low, medium, high) from a standard battery of reading tasks requiring varying levels of working memory. This is a challenging machine learning problem because of the inherent difficulty in defining/measuring cognitive load and due to intra-/inter-speaker differences in how their effects are manifested in behavioral cues. We experimented with a number of static and dynamic features extracted directly from the audio signal (prosodic, spectral, voice quality) and from automatic speech recognition hypotheses (lexical information, speaking rate). Our approach to classification addressed the wide variability and heterogeneity through speaker normalization and by adopting an i-vector framework that affords a systematic way to factorize the multiple sources of variability.

Index Terms: computational paralinguistics, behavioral signal processing (BSP), prosody, ASR, i-vector, cognitive load

1. Introduction

Cognitive load is related to the amount of information working memory can simultaneously hold and process. Automatically detecting an individual's cognitive load has many practical applications, e.g., in the monitoring of call center operators; these technologies could enable a reduction in stress, fatigue, and human error by identifying when there is information overload.

One promising and non-intrusive way to automatically differentiate individuals' level of cognitive load is through speech analysis. Previous studies have shown that variability and other statistics of prosodic cues (pitch/f_0, intensity, speaking rate) are correlated with cognitive load [1–4]. Specifically, "flattened" intonation was found to be a good indicator of high cognitive load in [2]. Spectral features, such as spectral peak prominence, spectral centroid frequency and amplitude, spectral energy spread, and formant frequencies (particularly low formants) and their trajectories, contain information regarding the cognitive load of a speaker [1, 2, 5, 6]. Voice source characteristics, including the variation of primary open quotient, normalized amplitude quotient, and primary speed quotient, are also effective cues [7]. Voice quality features (creakiness, harmonics-to-noise ratio), glottal flow shape, speech phase (group delay, FM parameter), and lexical/linguistic information (e.g., the duration and number of pauses and fillers) were also reported as important cues for detecting cognitive load level [1, 7, 8].

In this work, we apply emerging *behavioral signal processing* (BSP) methodologies to robustly classify speakers' level of cognitive load from an existing behavioral data corpus. BSP is centered on modeling societally-significant problems that are more abstract/subjective in nature, often characterizing (effects of) a person's internal state or being [9], e.g., detecting *blame* in married couples' interactions [10] and modeling therapists' *empathy* in drug addiction counseling [11]. Signal processing techniques are first used to extract relevant features from human behavioral signals (e.g., speech, language). Machine learning techniques are then used to map these features to the relevant higher-level descriptions. The specific methods employed in this paper are based on our previous INTERSPEECH Challenge work [12–14] and adapted to the specific computational paralinguistics problem of classifying cognitive load level.

Section 2 describes the corpus, and Section 3 discusses the acoustic features we analyzed in detail. Section 4 explains the speaker-normalized i-vector framework we implemented to account for the various sources of variability in modeling the expression of cognitive load in speech. We report our results and provide a discussion in Section 5, and we offer our conclusions and plans for future work in Section 6.

2. Corpus

We used the Cognitive Load with Speech and EGG (CSLE) database [1], which includes audio recordings of participants reading prompts with varying levels of cognitive load. As recommended by the INTERSPEECH 2014 Challenge organizers [15], we only analyzed the reading *span* sentence task [16] and the two variants of the Stroop test [17]: Stroop *time* pressure task and Stroop *dual* task. The database was comprised of 3 speaker-disjoint sets: train, development, and test. Speaker labels were only provided for the train (11 subjects) and development (7 subjects) sets. Three-class cognitive load labels were included as part of the data, representing the load each prompt was expected to place on the cognitive system (L_1: low, L_2: medium, L_3: high). Please see [1, 15] for more details.

As part of this work, we manually transcribed a portion of the train/development utterances at the word-level. Phonetic spellings of partial words (e.g., due to false starts) were also transcribed, along with instances of laughter. These enriched transcriptions were used to design an appropriately constrained automatic speech recognition (ASR) system.

3. Acoustic Feature Analysis

3.1. Speaker Normalization & Clustering

Our underlying hypothesis, supported by the literature reviewed in Sec. 1, is that the acoustics of the subjects' speech will vary across cognitive load levels. Importantly, cognitive load level will affect individuals' speech patterns in different ways. In

order to minimize the effect of inter-speaker variability on classification performance, speaker normalization techniques were employed throughout our analysis. As noted in Sec. 2, the test set did not have speaker labels, so we performed unsupervised speaker clustering on this data. First, we removed silence regions in each utterance using a statistical model-based voice activity detector (VAD) [18]. Then, a single Gaussian-based bottom-up agglomerative hierarchical clustering was performed in the linear predictive coding feature space [19]. The generalized likelihood ratio was used as the inter-cluster distance measure [20]. We correctly clustered 98.1% and 99.9% of the utterances in the train and development sets, respectively, suggesting the speaker clustering hypotheses on the test set were accurate.

3.2. Prosodic Features

We first examined the use of prosodic features (silence, f_0, intensity) because they are easily interpreted. We used the VAD (Sec. 3.1) to compute the mean and standard deviation (SD) of silence region durations for each utterance; we used Praat [21] to estimate f_0 and intensity and calculated the mean, SD, skewness, kurtosis, minimum, and maximum. In addition to these static features, we also modeled the dynamics of silence trends by plotting the duration of silence regions as a function of time; we then found a best fit line and used the slope and the mean square error (MSE) as additional utterance-level features. Similarly, we created a scatter plot of f_0 and intensity as a function of time and computed the slope and MSE of the best fit line.

Table 1 shows average prosodic signal statistics, computed across all utterances in the train and development sets, for each task and cognitive load level separately. We see from this table that the statistics of prosodic signals analyzed in this section displayed increasing or decreasing trends across cognitive load levels in the Stroop tasks more than in the span sentence task. Specifically, statistics of silence region durations were discriminative of the cognitive load levels for the Stroop dual task: as cognitive load level increased, silence regions tended to have shorter and more variable durations, and the MSE of the best fit line increased. f_0 statistics displayed trends across varying cognitive load levels for all three tasks: as the load level increased, mean f_0 increased for both Stroop tasks, and f_0 decreased at a decreasing rate for all three tasks (based on slope statistics). Intensity features were relevant for both Stroop tasks: as cognitive load level increased, subjects spoke louder in the Stroop time task, and variability in loudness increased in both Stroop tasks. Intensity decreased at a decreasing rate and MSE increased for both Stroop tasks as cognitive load level increased.

Sig.	Stat.	*span*			*time*			*dual*		
		L_1	L_2	L_3	L_1	L_2	L_3	L_1	L_2	L_3
	# utt.	378	378	594	54	54	54	54	54	54
Sil.	mean	.088	.086	.091	.268	.447	.413	**.619**	**.581**	**.573**
	SD	.067	.065	.076	.156	.310	.273	**.216**	**.297**	**.376**
	slope	.005	.002	.005	.002	.004	.001	**.001**	**-.001**	**-.002**
	MSE	.074	.073	.091	.162	.331	.265	**.210**	**.290**	**.371**
f_0 (Hz)	mean	138	136	135	**125**	**129**	**129**	**127**	**129**	**130**
	SD	25.0	23.5	23.8	17.8	18.0	17.6	17.9	18.9	17.8
	slope	**-15.5**	**-15.1**	**-15.1**	**-1.97**	**-.981**	**-.306**	**-.661**	**-.465**	**-.461**
	MSE	**21.7**	**20.6**	**20.5**	**19.4**	**20.4**	**20.4**	20.2	22.0	19.6
Int. (dB)	mean	72.4	72.2	72.3	**70.3**	**72.7**	**73.5**	70.9	72.5	72.4
	SD	5.13	5.06	5.18	**4.68**	**4.84**	**5.12**	**4.86**	**5.16**	**5.22**
	slope	-2.48	-2.44	-2.62	**-.380**	**-.110**	**-.048**	**-.101**	**-.055**	**-.028**
	MSE	4.61	4.56	4.59	**4.47**	**4.81**	**5.11**	**4.83**	**5.18**	**5.22**

Table 1: *Statistics of the prosodic signals, computed across all utterances in the train/development sets for each task and cognitive load level. Bold triplets highlight increasing or decreasing trends by load.*

3.3. Automatic Speech Recognition-Derived Features

To obtain word hypotheses and phone- and word-level boundaries, we performed ASR on each utterance. We used the Kaldi speech recognition toolkit [22] with triphone acoustic models trained on the Wall Street Journal database. Leveraging the transcriptions from Sec. 2, we trained task-specific language models by constraining the word hypotheses to those that occurred in the train/development sets. For the Stroop tasks, we used a unigram language model consisting of 10 color words (e.g., red, blue) and observed disfluencies. For the span sentence task, we trained a bigram language model using the transcribed text (674 words/disfluencies). We used the provided speaker labels for the train/development sets and the hypothesized speaker clusters from Sec. 3.1 for the test set to decode the utterances using Speaker Adaptive Training (SAT) [23]. We attained 20.4%, 11.3%, and 11.4% word error rates (WER) for the span sentence, Stroop time, and Stroop dual tasks, respectively. The disparity in WER between the tasks is most likely due to the significantly larger lexicon required for the span sentence task.

Initial analysis of the ASR hypotheses showed that the presence of disfluencies (e.g., fillers, laughter) was not correlated with cognitive load, so we did not examine these lexical features further. Instead, we concentrated on extracting speaking rate information from the ASR hypotheses. Specifically, we devised two systems that modeled the dynamics of phone-rate statistics and word durations, described in detail next.

3.3.1. Phone-rate statistics

We defined average phone speaking rate over a window as the number of non-silence phones in the window normalized by its duration. For each utterance, we evaluated this average phone rate measure at various time resolution scales. To evaluate the measure at scale i, the utterance was chopped into i equal-sized segments for which each the average phone rate was calculated and subsequently stacked in an i-dimensional vector. We then concatenated the vectors for scales 1 to N to obtain the final feature representation, with dimension $\frac{N(N+1)}{2}$. This feature characterizes the utterance in terms of the progression of phone rates across time in a multi-resolution fashion. We chose $N = 7$ for the Stroop tasks, and we chose $N = 5$ for the span sentence task, since utterance durations are much shorter. The feature was then normalized by task and speaker for each utterance. To obtain baseline results with this feature, we trained a support vector machine (SVM) on the train set, for each task separately, and applied it to the development set. See Sec. 5 for results.

3.3.2. Dynamics of word durations

We represented each utterance by the corresponding sequence of word durations, in seconds, treating inter-word silence as a word. Since the durations vary by speaker, we normalized the duration of each word by its corresponding mean and SD for each speaker. We also replaced non-color words (e.g., fillers) with a high positive SD of +3. Since the Stroop tasks have such a small lexicon, there is enough data available for each word to collect duration statistics for normalization. However, this is not the case for the span sentence task, where many words are only spoken once by a speaker. As a result, we only applied this analysis to the Stroop tasks.

To capture the dynamics of this normalized word-duration sequence, we trained a Hidden Markov Model (HMM) for each cognitive load level on the train set. Each HMM had a fully-connected three-state topology; the state means were initialized

to -2, 0, and 2, corresponding roughly to word durations much smaller than, equal to, and much larger than the mean, respectively. We only discuss the learned parameters for the Stroop *dual* task because of space constraints and because this methodology performed better on this task; please see Sec. 5. The state means are denoted as $\boldsymbol{\mu_{L_i}}$ and transition matrices as T_{L_i}:

$$\boldsymbol{\mu_{L_1}} = \begin{bmatrix} -0.88 \\ 0.04 \\ 0.75 \end{bmatrix}, \quad \boldsymbol{\mu_{L_2}} = \begin{bmatrix} -0.98 \\ 0.04 \\ 1.49 \end{bmatrix}, \quad \boldsymbol{\mu_{L_3}} = \begin{bmatrix} -1.15 \\ -0.02 \\ 1.73 \end{bmatrix} \quad (1)$$

$$T_{L_1} = \begin{bmatrix} .03 & .13 & .84 \\ .08 & .89 & .03 \\ .65 & .32 & .03 \end{bmatrix}, T_{L_2} = \begin{bmatrix} .34 & .25 & .41 \\ .03 & .83 & .14 \\ .72 & .15 & .12 \end{bmatrix}, T_{L_3} = \begin{bmatrix} .42 & .20 & .38 \\ .02 & .80 & .18 \\ .49 & .34 & .18 \end{bmatrix} \quad (2)$$

As shown in (1), extreme deviations from the mean were rare for L_1, and the deviations increased proportionally with the cognitive load level. We infer from (2) that there is greater volatility for the higher load levels, based on the findings that with increasing cognitive load, there was: decreased probability of state 2 self-transitions; increased probability of state 1 and 3 self-transitions; and increased probability of transitions into and out of state 2. We evaluated probabilities of each observation sequence in the development/test sets for the three HMMs and classified the utterance to the cognitive load level with the maximum probability; please see Sec. 5 for results.

4. i-Vector Modeling Framework

The proposed system for cognitive load level classification exploits the concept of *total variability* or *i-vector* modeling, originally proposed in [24] and motivated by Joint Factor Analysis (JFA) [25,26]. Total variability modeling provides an improved accuracy and reduced computational complexity compared to factor analysis by estimating a single low dimensional subspace, named the identity or *i-vector* space, in which all variability is modeled together. In this work, we attempt to model the speaker-specific variability across the cognitive load levels: L_1, L_2, and L_3. To compensate for the undesired variability due to utterance- and speaker-dependent factors, further variability compensation methods are applied within the i-vector space.

4.1. Feature Extraction and Background Modeling

In addition to the features explored in Sec. 3, we also extracted four diverse acoustic feature sets to train task-specific i-vector systems:

- *PLP*: Perceptual Linear Prediction (PLP) coefficients [27]
- *GBF*: Gabor features which capture the acoustic modulations patterns in speech [28]
- *ISCOMP11*: statistics of frame-level openSMILE [29] spectral, voice quality, and prosodic features [30]
- *FuSS*: Fused Speech Stream features obtained by combining the four speech streams of [31] into a single feature vector. Each of these streams models a different aspect of the human voice: (i) spectral shape, (ii) spectro-temporal modulations, (iii) periodicity structure due to the presence of pitch harmonics, and (iv) the long-term spectral variability profile.

All the above features are mean- and variance-normalized on a per speaker basis, deploying the speaker clustering strategy of Sec. 3.1. For each feature representation, a task-specific Universal Background Model (UBM) was trained using all available training and development set data.

4.2. Total Variability Modeling

The application of i-vector modeling for utterance-level classification of cognitive load implies representing each utterance j of speaker s as a supervector $\mathbf{M}_j^s$ [24]:

$$\mathbf{M}_j^s = \mathbf{m} + \mathbf{T}\mathbf{w}_j^s \quad (3)$$

where $\mathbf{m}$ is a load-, speaker-, and utterance-independent supervector constructed from stacking the Gaussian mean vectors of all UBM components. The total variability matrix $\mathbf{T}$ is of low rank and obtained by task-specific factor analysis training. The i-vector of the speaker's utterance is then given by a normally distributed vector $\mathbf{w}_j^s$ containing the corresponding total factors [24]. The rank of $\mathbf{T}$ defines the dimensionality of the i-vectors and is tuned for each task (*span*: 200; *time*, *dual*: 75). The i-vector model of (3) is extended to the simplified framework of [32,33] to reduce computational complexity and is iteratively trained by the Expectation Maximization method; see [25, 34] for more details.

4.3. Speaker-Dependent i-Vector Normalization

We can assume that the extracted i-vectors are highly biased to the task-specific behavior of speakers, regardless of the cognitive load level of the task, and that the acoustics of the speaker varies with respect to this bias when exposed to a different load level. Since the system is trained on multiple speakers, the bias term acts as a noise factor in the i-vector space and hence needs to be factored out. Therefore, we assume that $\mathbf{w}_j^s$ in (3) can be written as:

$$\mathbf{w}_j^s = \mathbf{b}^s + \mathbf{z}_j^s \quad (4)$$

where $\mathbf{b}^s$ and $\mathbf{z}_j^s$ respectively correspond to the speaker-dependent bias and a term that captures the residual variability in speaking style due to the task's load level. The residual term is obtained by first estimating the bias $\mathbf{b}^s$ as the i-vector mean over all load levels per speaker (using the supplied speaker labels for the train/development sets and the Sec. 3.1 speaker clustering results for the test set) and subsequently subtracting this estimate from $\mathbf{w}_j^s$. What remains is a speaker-independent term that better models the load level variability and hence will serve as the input features on which a classifier is learned. Classification is done by training an SVM with polynomial kernel (fifth order) on the residual term of all training utterances using the load levels as class labels. For each feature representation of Sec. 4.1, a system was trained for the development and test set (on the training set and training plus development sets, respectively), where the number of UBM components and i-vector dimensions are optimized per task using a leave-one-speaker-out cross-validation strategy.

5. Results & Discussion

As set by the INTERSPEECH 2014 Challenge organizers [15], we used *Unweighted Average Recall* (UAR) as the evaluation metric for all systems, defined as the unweighted (by number of utterances in each class) mean of the percentage correctly classified in the diagonal of the confusion matrix. Please see Table 2 for the performance of the various proposed automatic systems; note that the provided baseline system consists of a linear SVM classifier trained on 6373 functionals of spectral/prosodic/voice quality low-level descriptors [15].

The numbers reported for the i-vector systems in Table 2 were obtained through mean-variance normalization on

System	Feature(s)	*span*	*time*	*dual*	Total
	Chance	33.3	33.3	33.3	**33.3**
SVM	*Baseline* [15]	61.2	74.6	63.5	**63.2**
SVM	*Prosody* (Sec. 3.2)	41.7	88.9	77.8	**49.8**
SVM	*Phone-Rate* (Sec. 3.3.1)	48.5	54.0	76.2	**52.7**
HMM	*Word Durations* (Sec. 3.3.2)	—	57.1	77.8	**—**
i-vector	1: *PLP* (Sec. 4.1)	62.9	71.4	65.1	**63.9**
i-vector	2: *GBF* (Sec. 4.1)	72.0	84.1	73.0	**73.3**
i-vector	3: *ISCOMP11* (Sec. 4.1)	69.9	73.0	71.4	**70.3**
i-vector	4: *FuSS* (Sec. 4.1)	75.1	82.5	81.0	**76.4**
i-vector	*Fusion*: 1+2+3+4	76.0	84.1	82.5	**77.5**

Table 2: *Unweighted Average Recall (%) results on the development set for each task: reading span sentence, Stroop time pressure, Stroop dual.*

a speaker basis. The beneficial effect of speaker compensation is shown in Table 4 for the i-vector system trained on the *FuSS* feature only (because of space constraints). The application of mean normalization on the i-vector space results in an increase in performance for all tasks. The addition of variance normalization boosts performance for both Stroop tasks. Results not reported here show that the benefit of variability compensation in this task is consistent for all feature representations and more effective than, e.g., Within-Class Covariance Normalization (WCCN), and thus can be considered an original contribution of this work. These speaker-normalized i-vector systems in Table 2 (1-4) were also augmented with the prosody (Sec. 3.2) and phone-rate (Sec. 3.3.1) features. Running the SVM classifier on these augmented vectors gave us a slight performance improvement, as shown in Table 4 for the *FuSS* features.

As can be seen in Table 2, in general, performance was worse for the span sentence task compared to the Stroop tasks. This is most likely due to the fact that the utterances in the span sentence task were significantly shorter in duration and phonetically more challenging, due to the limited vocabulary size of the Stroop tasks. Therefore, there is an additional source of *lexical* variability present in the span sentence task.

The features analyzed in Sec. 3 performed well for the Stroop dual task, and the prosodic features analyzed in Sec. 3.2 also excelled in the Stroop time task. However, these systems did not generalize well to the span sentence task. This disparity in performance across tasks is likely due to a number of factors: 1) the features explored in Sec. 3 were specifically targeting prosodic cues (speech/silence, f_0, intensity, speaking rate) that we observed in the train data to be perceptually relevant for the Stroop tasks; 2) there are fewer utterances in the Stroop tasks (Table 1), so the Sec. 3 systems, which have significantly lower dimensional feature spaces compared to the baseline system, are less susceptible to the *curse of dimensionality* problem; and 3) the Stroop tasks do not have continuous speech, so it is easier to track changes in speech cues within an utterance.

Our best performing system overall (the final row in Table 2) was obtained by exploiting the complementary information of i-vector systems 1-4 through linear fusion of the output probabilities. Please see Table 3 for a confusion matrix of the best proposed automatic system. As was expected, the most difficult cognitive load level to classify was L_2 for all three tasks. This more ambiguous "middle" level is more easily confused with L_1 and L_3 because it is inherently perceptually closer, due

System	*span*			*time*			*dual*		
	L_1	L_2	L_3	L_1	L_2	L_3	L_1	L_2	L_3
L_1	**75.5**	9.5	6.9	**90.5**	0.0	0.0	**85.7**	19.0	0.0
L_2	10.9	**71.4**	12.1	9.5	**76.2**	14.3	9.5	**76.2**	14.3
L_3	13.6	19.0	**81.0**	0.0	23.8	**85.7**	4.8	4.8	**85.7**

Table 3: *Confusion matrices (%) across the 3 cognitive load levels for the best overall automatic system on the development set for each task.*

Feature(s)	Normalization	*span*	*time*	*dual*	Total
FuSS	none	74.7	66.7	63.0	**72.8**
FuSS	mean	75.0	78.7	77.3	**75.6**
FuSS	mean-variance	74.7	81.7	80.3	**75.9**
FuSS+Prosody+Phone-Rate	mean-variance	75.1	82.5	81.0	**76.4**

Table 4: *UAR (%) results for the i-vector system trained with FuSS features using different speaker normalization methods for each task.*

to the ordinal nature of the class labels. Future work will look to explicitly model this ordinality constraint.

Finally, as part of this work, we tested our best performing system on the distinct but overlapping problem domain of detecting *physical load*, which was also part of the INTERSPEECH 2014 Challenge [15]. The purpose of this experiment was to see if the methodologies we developed to distinguish varying levels of cognitive load generalized to the similar problem of separating speech produced under low and high levels of physical load. We used the Munich Bio-voice Corpus (MBC) [35], which includes recordings of subjects reading a passage in a baseline resting state (*low* physical load) and after exercising (*high* physical load). The corpus was set up in a similar manner to the reading span sentence task in the CLSE database (Sec. 2), with speaker-disjoint train, development, and test sets. Table 5 shows that our best proposed system was able to generalize to new speakers and related machine learning problems, outperforming the baseline system on the development and test sets for both INTERSPEECH 2014 Sub-Challenges: detecting *cognitive* and *physical* load. This suggests that the i-vector framework proposed in this work may be appropriate for other important computational paralinguistics problems, an area of future work.

6. Conclusion & Future Work

We explored the use of multiple acoustic features to classify the cognitive load level of speakers' utterances as low, medium, or high. Through feature-level fusion of these features within a novel speaker-normalized i-vector framework, we were able to beat the baseline SVM approach. This work is promising for real-world applications, such as monitoring of subjects who perform cognitively demanding tasks (e.g., call center operators).

Future work will look to improve the classification accuracy on the more ambiguous L_2 utterances, potentially through multi-stage hierarchical methodologies or by exploiting the ordinal nature of the labels, e.g., by using ordinal logistic regression techniques [36]. We also plan to experiment with other fusion methodologies at the feature-, score-, and classifier-level. Finally, we will continue to apply the proposed automatic systems to related problem domains in behavioral signal processing (BSP), such as physical load level classification.

7. Acknowledgements

This research was supported by NSF, DARPA, and NIH. We would like to acknowledge the work completed by Daniel Bone following the initial submission of the paper. Special thanks to the Interspeech Challenge organizers and to Mary Francis for her devotion and help in all SAIL research efforts.

System	*Cognitive Load*		*Physical Load*	
	Dev	Test	Dev	Test
Chance	33.3	**33.3**	50.0	**50.0**
Baseline [15]	63.2	**61.6**	67.2	**71.9**
Fusion: 1+2+3+4	77.5	**68.9**	71.8	**73.9**

Table 5: *Unweighted Average Recall (%) results on the Development (Dev) and Test sets for both INTERSPEECH 2014 Sub-Challenges [15].*

8. References

[1] T. F. Yap, "Speech production under cognitive load: Effects and classification," Ph.D. dissertation, The University of New South Wales, 2012.

[2] P. N. Le and E. Choi, "The use of spectral information in the development of novel techniques for speech-based cognitive load classification," Ph.D. dissertation, School of Electrical Engineering and Telecommunications, The University of New South Wales, Australia, 2012.

[3] M. Charfuelan and G.-J. Kruijff, "Analysis of speech under stress and cognitive load in USAR operations," in *Natural Interaction with Robots, Knowbots and Smartphones*, 2014, pp. 275–281.

[4] 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.

[5] P. N. Le, E. Ambikairajah, J. Epps, V. Sethu, and E. H. C. Choi, "Investigation of spectral centroid features for cognitive load classification," *Speech Communication*, vol. 53, no. 4, pp. 540–551, 2011.

[6] H. Boril, S. O. Sadjadi, T. Kleinschmidt, and J. H. L. Hansen, "Analysis and detection of cognitive load and frustration in drivers' speech," in *Proc. Interspeech*, 2010.

[7] T. F. Yap, J. Epps, E. H. Choi, and E. Ambikairajah, "Glottal features for speech-based cognitive load classification," in *Proc. IEEE ICASSP*, 2010.

[8] A. Jameson, J. Kiefer, C. Müller, B. Großmann-Hutter, F. Wittig, and R. Rummer, "Assessment of a user's time pressure and cognitive load on the basis of features of speech," in *Resource-Adaptive Cognitive Processes*, 2010, pp. 171–204.

[9] S. S. Narayanan and P. G. Georgiou, "Behavioral Signal Processing: Deriving human behavioral informatics from speech and language," *Proc. of IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[10] M. P. Black, A. Katsamanis, B. Baucom, C.-C. Lee, A. Lammert, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features," *Speech Communication*, vol. 55, no. 1, pp. 1–21, 2013.

[11] B. Xiao, P. G. Georgiou, Z. E. Imel, D. Atkins, and S. S. Narayanan, "Modeling therapist empathy and vocal entrainment in drug addition counseling," in *Proc. Interspeech*, 2013.

[12] D. Bone, M. Li, M. P. Black, and S. S. Narayanan, "Intoxicated speech detection: A fusion framework with speaker-normalized hierarchical functionals and GMM supervectors," *Computer Speech and Language*, vol. 28, no. 2, pp. 375–391, 2014.

[13] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, "Intelligibility classification of pathological speech using fusion of multiple high level descriptors," in *Proc. Interspeech*, 2012.

[14] R. Gupta, K. Audhkhasi, S. Lee, and S. S. Narayanan, "Speech paralinguistic event detection using probabilistic time-series smoothing and masking," in *Proc. Interspeech*, 2013.

[15] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPECH 2014 computational paralinguistics challenge: Cognitive & physical load," in *Proc. Interspeech*, Singapore, Singapore, 2014.

[16] M. Daneman and P. A. Carpenter, "Individual differences in working memory and reading," *Journal of Verbal Learning and Verbal Behavior*, vol. 19, no. 4, pp. 450–466, 2005.

[17] J. R. Stroop, "Studies of interference in serial verbal reactions," *Journal of Experimental Psychology*, vol. 18, no. 6, p. 643, 1935.

[18] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan 1999.

[19] W. Wang, P. Lu, and Y. Yan, "An improved hierarchical speaker clustering," *ACTA ACUSTICA*, 2008.

[20] K. J. Han, S. Kim, and S. S. Narayanan, "Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1590–1601, 2008.

[21] P. P. G. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[23] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP*, 1996.

[24] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[25] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[26] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[27] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[28] M. Kleinschmidt, "Spectro-temporal Gabor features as a front end for ASR," in *Proc. Forum Acusticum Sevilla*, 2002.

[29] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE - The Munich versatile and fast open-source audio feature extractor," in *ACM Multimedia*, Firenze, Italy, 2010, pp. 1459–1462.

[30] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Proc. Interspeech*, 2011.

[31] M. Van Segbroeck, A. Tsiartas, and S. S. Narayanan, "A robust frontend for VAD: Exploiting contextual, discriminative and spectral cues of human voice," in *Proc. Interspeech*, 2013.

[32] M. Li, A. Tsiartas, M. Van Segbroeck, and S. S. Narayanan, "Speaker verification using simplified and supervised i-vector modeling," in *Proc. IEEE ICASSP*, 2013.

[33] M. Li and S. S. Narayanan, "Simplified supervised i-vector modeling with application to robust and efficient language identification and speaker verification," *Computer Speech and Language*, 2014.

[34] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction." in *Proc. Interspeech*, 2011.

[35] B. Schuller, F. Friedmann, and F. Eyben, "The Munich Biovoice Corpus: Effects of physical exercising, heart rate, and skin conductance on human speech production," in *Language Resources and Evaluation Conference*, 2014.

[36] V. Rozgić, B. Xiao, A. Katsamanis, B. Baucom, P. G. Georgiou, and S. S. Narayanan, "Estimation of ordinal approach-avoidance labels in dyadic interactions: Ordinal logistic regression approach," in *Proc. IEEE ICASSP*, 2011.