The UNSW Submission to INTERSPEECH 2014 ComParE Cognitive Load Challenge

Jia Min Karen Kua[1], Vidhyasaharan Sethu[1], Phu Le[1] and Eliathamby Ambikairajah[1,2]

[1]The School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney NSW 2052, Australia
[2]National ICT Australia (NICTA), Australian Technology Park, Eveleigh 1430, Australia

j.kua@unsw.edu.au, v.sethu@unsw.edu.au, phule@unsw.edu.au, ambi@ee.unsw.edu.au

Abstract

Speech based cognitive load estimation is a new field of research. Due to this relative 'lack of maturity', a single best approach to building cognitive load estimation systems has not been established yet. The primary aim of this submission is to report the performance of various basic utterance level classification frameworks developed using important elements of state-of-the-art speaker recognition systems. This may lead to a suitable basis for future cognitive load estimation systems. As a consequence of being a part of a challenge, it is expected that these frameworks will be compared to a much larger number of alternative approaches than what would otherwise be possible. In keeping with this focused aim, the GMM supervector approaches along with some variants are utilised. The systems outlined in this paper include a frame-level MFCC-GMM system along with utterance level GMM-supervector-SVM, GMM-ivector-SVM and GMM-JFA-SVM systems. The best combined system has an accuracy (UAR) of 66.6% as evaluated on the challenge development set and 63.7% as evaluated on the test set.

Index Terms: cognitive load estimation, GMM supervectors, support vector machines

1. Introduction

The cognitive load of a person refers to the amount of mental demand imposed on the person when performing a particular task and has been closely associated with the limitations of human working memory [1, 2]. Research on cognitive load has shown that when a user is performing a task, performance will degrade if the load level is too low or too high [1]. In some cases however, it may be possible to adjust the workload of the user and if the cognitive load level of the user can be estimated or classified along an ordinal scale, it may be possible to tailor these adjustments to the workload such that productivity can be improved.

Over the last two decades, a number of techniques based on physiological measures [3, 4], behavioural measures [5, 6] and performance measures [1, 7] have been proposed for measuring cognitive load level. Among these, behavioural measures based on speech have been recognised as a particularly good choice since speech data exists in many real-life tasks (e.g. telephone conversation, voice control) and can be easily collected in non-intrusive and inexpensive ways.

In speech-based cognitive load classification systems, cepstral features extracted on a frame-by-frame basis are by far the most popular features [8-10]. Moreover, MFCCs can be considered as de-facto reference features for cognitive load classification, since most current studies compare the performance of newly proposed features to a baseline system involving MFCC features. Apart from cepstral features, frequency-based features have also been recognised to be effective for cognitive load classification [8, 10, 11]. However, speech based cognitive load estimation is still a relatively new field of research and no single system has been unequivocally established as the best approach to adopt. Given this, the systems explored in this paper are relatively basic systems with little or no specialised channel and/or speaker normalisation. However, the approaches adopted in these systems form the core of recent and past standard speaker recognition systems [12-15].

2. Database

In the Cognitive Load Sub-Challenge, speech and electroglottograph data (CLSE database) are provided as part of the challenge for identifying the cognitive load level of a subject [16]. In this paper, only the speech data is used in all experiments. The speech was recorded in Australia using a close-talk microphone sampled at 16kHz from 26 native Australian English speakers while undertaking four different speaking tasks (readingspanSentence, readingspanLetter, strooptimepressure and stroopdualtask) [16]. There are three cognitive load levels to be distinguished: low (L1), medium (L2) and high (L3).

The database is divided into speaker disjoint training, development and test sets with all 4 different tasks in each. Furthermore, additional neutral speech is provided for training suitable background models. While data is available for 4 tasks, speech recorded from the readingspanLetter task is of very short duration and is not included in the designated test set [16]. Consequently in all the experimental work reported in this paper, data corresponding to the readingspanLetter task was left out of the training and development sets as well. For further details on partitioning of the database the reader is referred to Schuller et al [16].

3. System Descriptions

A number of different systems were evaluated on the development set prior to selecting the final systems for the challenge. These systems are described in this section.

3.1. Frame-level MFCC-GMM system

The basic frame-level cepstral cognitive load classification system consists of a MFCC+ΔMFCC+ΔΔMFCC (12+12+12 dimensions) front-end that operates with 20ms frames with 10ms overlap using a Hamming window. The back-end is based on Gaussian mixture models (GMMs) with a GMM, $\mathcal{G}_k$, trained for each class (k) via ML (maximum likelihood) estimation. The basic system does not make use of any normalisation or adaptation techniques. For each utterance, $\mathfrak{U}$, the estimated cognitive load ($\bar{k}$) is given by eqn (2).

$$\Lambda_K(\mathfrak{U}) = \sum_{t=1}^{N_T} \log P(\mathbf{x}_t | \mathcal{G}_k) \tag{1}$$

$$\bar{k}(\mathfrak{U}) = \arg\max_k \Lambda_k(\mathfrak{U}) \tag{2}$$

Where, $\mathbf{x}_t$ is the feature vector corresponding to the t^{th} frame of the utterance $\mathfrak{U}$, $P(\cdot\,|\,\cdot)$ denotes conditional probability, N_T is the total number of frames per utterance and

$\mathcal{G}_k$ is the GMM trained on the data corresponding to cognitive load level k.

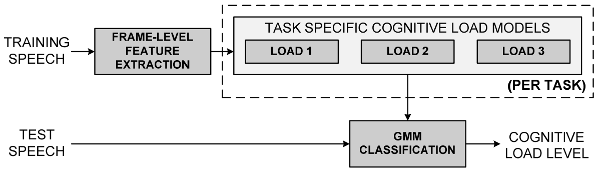

Figure 1: Overview of basic frame-level MFCC-GMM system

The basic frame level MFCC based system only models frame based features and does not take into account any long term trends beyond those captured by the $\Delta\Delta$ components.

3.2. Utterance-level GMM supervector systems

In general there are two broad approaches for taking into account the long term information not captured by frame based features. These include (a) the use of appropriate back-ends, such as hidden Markov models, which model temporal patterns of short term features; and (b) the estimation of utterance-level representations of speech, typically estimated from frame-based features. A commonly utilised utterance-level representation, almost universally adopted in speaker recognition systems [13], is the Gaussian mixture model (GMM) supervector.

In general, GMM supervectors are estimated by first training a Gaussian mixture model, referred to as a universal background model (UBM), on generic speech representative of the style that is expected to be confronted by the cognitive load estimation system. This UBM, $\mathcal{G}_{UBM}$, is then adapted (via MAP adaptation) towards each utterance to obtain an estimate of a model of the feature distributions corresponding to them, $\mathcal{G}_{\mathfrak{U}}$. A GMM supervector is a vectorial representation of the parameters of these models. In the common case where only the UBM means are adapted via MAP adaptation, the GMM supervector is composed by concatenating the means of each mixture component of the adapted GMM and is given by $\boldsymbol{M}(\mathcal{G}_{\mathfrak{U}}) = [\boldsymbol{\mu}_1^T\ \boldsymbol{\mu}_2^T \dots\ \boldsymbol{\mu}_N^T]^T$, where $\boldsymbol{\mu}_i \in \mathbb{R}^D$ is the mean of the i-th Gaussian component.

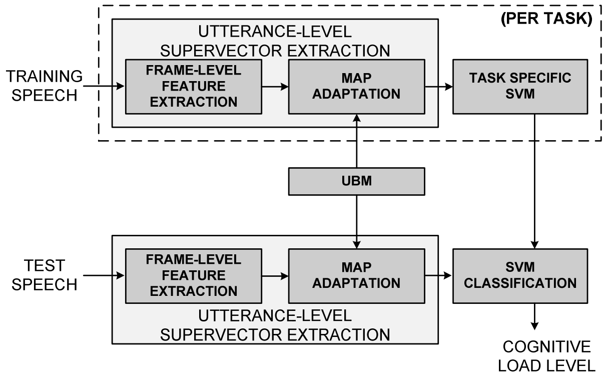
(PER TASK)
UTTERANCE-LEVEL SUPERVECTOR EXTRACTION
TRAINING SPEECH
FRAME-LEVEL FEATURE EXTRACTION
MAP ADAPTATION
TASK SPECIFIC SVM
UBM
TEST SPEECH
FRAME-LEVEL FEATURE EXTRACTION
MAP ADAPTATION
SVM CLASSIFICATION
UTTERANCE-LEVEL SUPERVECTOR EXTRACTION
COGNITIVE LOAD LEVEL

Figure 2: Overview of utterance-level GMM supervector SVM system

Gaussian mixture models are parametric models of feature distributions. Consequently GMM supervectors extracted from each utterance are high dimensional vectorial representations of the distributions of frame level features obtained from that utterance. This paper includes results from two GMM supervector sub-systems that model the distributions in different feature spaces. Specifically, an MFCC based feature space and a spectral centroid based feature space, both of which have been shown to be effective in modelling cognitive load information [8].

Following the estimation of supervectors, a support vector machine (SVM) based back-end is trained to classify them into one of the three cognitive load levels. A separate SVM is trained for each of the three tasks (readingspanSentence, strooptimepressure and stroopdualtask) and all testing is done in a task dependent manner.

3.2.1. MFCC+SDC sub-system

The front-end of the MFCC sub-subsystem extracts standard Mel frequency cepstral coefficients (12 dimensions including C_0) appended with shifted delta coefficients (SDCs) [17] to capture the long-term information of the features. The configuration of the SDCs is given by three parameters: D_F, P, and K. D_F is the number of frames used to compute each delta, P is the number of frames between consecutive deltas and K is the total number of shifted delta values concatenated together to form the shifted delta feature as given by eqn (3). For each of the feature streams, the shifted delta feature vector at time n is given by the concatenation of the $\Delta C_i(n,m)$ for $0 \leq m \leq K$. The parameters D_F,P and K are set to 1,3 and 7 [8].

$$\Delta C_i(n,m) = \frac{\sum_{d=-D_F}^{D_F} dC_i(n+mP+d)}{\sum_{d=-D_F}^{D_F} d^2} \tag{3}$$

3.2.2. SCF sub-system

The spectral centroid frequency (SCF) is an estimate of the 'centre of gravity' of the spectrum of speech. Typically the speech signal is split into a number of auditory subbands and the SCF is computed within each band. Originally proposed as a feature for speech recognition systems [18], it has been reported that SCF is a formant-like feature, as it provides the approximate location of the formant frequencies in the subbands [18, 19].

The stages involved in computing the spectral centroid frequency feature vector is schematically depicted in Figure 3.The spectral centroid frequency is extracted from framed speech segments of 20ms as follows. Let $s[n]$, where $n \in [0, F-1]$, represent a speech frame (of length F) in the time domain and let $S[f]$ represent the spectrum of this frame. Then, $S[f]$ is divided into N_S subbands by using a series of Gabor filters [20] whose frequency responses are $W_m[f]$, where $m \in [1, N_S]$. The number of subbands, $N_S = 21$ in this paper.

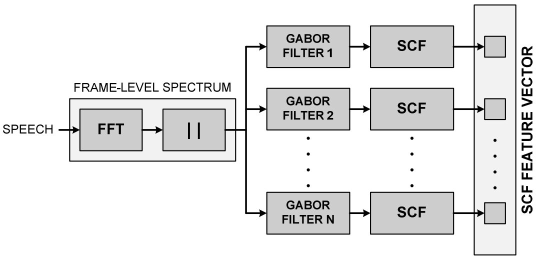

Figure 3 Spectral centroid frequency extraction

If the lowest frequency of the m^{th} subband is denoted by l_m and the highest frequency by u_m. The SCF of the m^{th} subband (SCF_m) is computed as the weighted average frequency, where the weights are the normalised energy of each frequency component, as given by eqn (4).

$$SCF_m = \frac{\sum_{\omega=l_m}^{u_m} \omega|S[\omega]W_m[\omega]|}{\sum_{\omega=l_m}^{u_m} |S[\omega]W_m[\omega]|} \tag{4}$$

The final SCF feature vector of each frame is obtained by concatenating all the SCF_m. Previous work has shown that SCF and formant feature complement the basic MFCC features in cognitive load classification system [8, 11].

3.3. Utterance-level i-vector based system

The use of GMM supervectors allows for the application of a number of linear vector space operations but may be held back by the inherent high dimensionality of supervectors. For instance, the GMM supervectors from the MFCC+SDC sub-system outlined in section 3.2.1 have a dimensionality of 24,576 since the system uses a 256 mixture UBM modelling a 96 dimensional feature space (MFCC+SDC). The i-vector space is a low dimensional subspace onto which supervectors are mapped via a linear transformation while retaining most of the variability (useful information) present in the supervector space and has been used extensively in speaker recognition. In the context of cognitive load estimation systems outlined in this paper, we expect the UBM to roughly model the phonetic structure of the acoustic feature space represented by MFCCs+SDCs (section 3.2.1) and the spectral centroid frequency features (section 3.2.2). Consequently the supervectors and the i-vectors would capture variations from this structure due to non-phonetic factors including speaker characteristics and cognitive load level. The i-vector model is given by:

$$\boldsymbol{M}(\mathcal{G}_{\mathfrak{U}}) = \boldsymbol{M}(\mathcal{G}_{UBM}) + \mathbf{T}\boldsymbol{w}_{\mathfrak{U}} \quad (5)$$

where $\mathcal{G}_{UBM}$ denotes the UBM, $\mathcal{G}_{\mathfrak{U}}$ denotes the GMM obtained via MAP adaptation of the UBM to approximate the distribution of features estimated from the utterance $\mathfrak{U}$, $\boldsymbol{w}_{\mathfrak{U}}$ is the i-vector corresponding to $\mathfrak{U}$ and **T** is the projection matrix [14].

The i-vector model is a factor analysis method and the **T** matrix is estimated from training data. In general, results from speaker recognition show that the more training data used in the training of the **T** space, the more accurate the final system [14]. In the system reported in this paper, the projection matrix was estimated from the training set, treating each cognitive load level under each task as distinct sessions. i.e., the **T** matrix was trained to capture the variability between different cognitive load levels for different tasks.

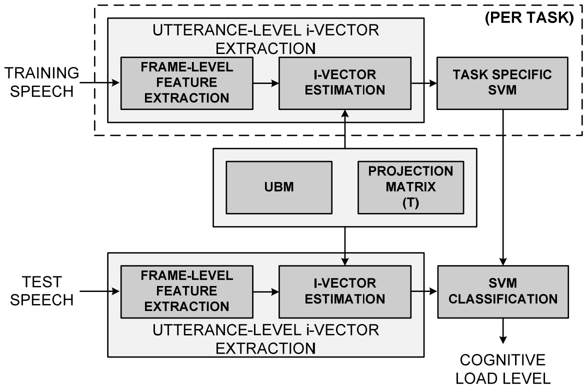

Figure 4: Overview of utterance level i-vector SVM system

Similar to the supervector system (section 3.2), following the extraction of i-vectors, a task dependent SVM back-end was used as a classifier to determine cognitive load level. Only the MFCC+SDC front-end was used in the i-vector extraction system.

3.4. Utterance-level JFA based system

Joint factor analysis (JFA) is factor analysis model similar to the i-vector model and has been widely used in speaker verification systems [15]. In the GMM-JFA paradigm, GMM supervectors are decomposed into a linear combination of factors that model target class variability (cognitive load factors in this case) and a second set of factors that model nuisance variability (speaker variability in this case). Mathematically, the model is given by:

$$\boldsymbol{M}(\mathcal{G}_{\mathfrak{U}}) = \boldsymbol{M}(\mathcal{G}_{UBM}) + \mathbf{V}y_{\mathfrak{U}} + \mathbf{U}x_{\mathfrak{U}} + \mathbf{D}\boldsymbol{z} \quad (6)$$

where $\boldsymbol{M}(\mathcal{G}_{\mathfrak{U}})$ is the GMM supervector corresponding to the GMM $\mathcal{G}_{\mathfrak{U}}$ obtained from utterance $\mathfrak{U}$, $\boldsymbol{M}(\mathcal{G}_{UBM})$ is the UBM supervector, $\mathbf{V} \in \mathbb{R}^{ND \times N_V}$ is a matrix of 'eigenloads' (analogous to eigenvoices), $\mathbf{U} \in \mathbb{R}^{ND \times N_U}$ is a matrix of 'eigenspeakers' (analogous to eigenchannels), $\mathbf{D} \in \mathbb{R}^{ND \times ND}$ is a diagonal matrix, $y_{\mathfrak{U}} \in \mathbb{R}^{N_V}$ is a vector of cognitive load factors, $x_{\mathfrak{U}}$ is a vector of speaker factors, $z \in \mathbb{R}^{ND}$ is a random vector and $\mathbf{D}\boldsymbol{z}$ represents variability in the supervector space not in the span of the eigenload or the eigenspeaker matrices.

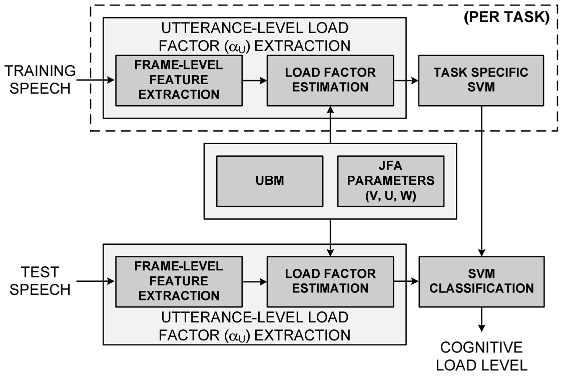
(PER TASK)
UTTERANCE-LEVEL LOAD FACTOR (α_U) EXTRACTION
TRAINING SPEECH
FRAME-LEVEL FEATURE EXTRACTION
LOAD FACTOR ESTIMATION
TASK SPECIFIC SVM
UBM
JFA PARAMETERS (V, U, W)
TEST SPEECH
FRAME-LEVEL FEATURE EXTRACTION
LOAD FACTOR ESTIMATION
SVM CLASSIFICATION
UTTERANCE-LEVEL LOAD FACTOR (α_U) EXTRACTION
COGNITIVE LOAD LEVEL

Figure 5: Overview of utterance level JFA SVM system

The utterance-level GMM-JFA system also utilises a task dependent SVM back-end.

3.5. Utterance-level Baseline Functional features

The openSMILE base feature set provided by the challenge organisers contains a number of common acoustic low-level descriptors (LLDs) [16]. It includes MFCCs, energy, jitter, shimmer and fundamental frequency (f_o). The final feature is based on functional (e.g. mean, standard deviation) of the LLD contours. This paper includes some results obtained by combining these baseline features with other.

3.6. Combining Utterance-level systems

In addition to comparing the individual systems outlined in this paper, the performance of combined systems were also estimated. The frame-level MFCC system was not used in any of the combined systems. All combinations of utterance-level systems were implemented by concatenation of the utterance-level representations (supervectors, i-vectors, functional features) to form a higher dimensional representation prior to using a SVM back-end. i.e., SVM back-end are trained and evaluated on concatenated utterance-level vectors, $\widehat{\Phi}$, which are given by:

$$\widehat{\Phi}(\mathfrak{U}) = [\Phi_1^T(\mathfrak{U}), \Phi_2^T(\mathfrak{U}), \ldots, \Phi_S^T(\mathfrak{U})]^T \quad (7)$$

Where $\Phi_x(\mathfrak{U}) \in \mathcal{R}^{D_x \times 1}$ denotes the D_x-dimensional vectorial representation of utterance $\mathfrak{U}$ corresponding to the x^{th} system that is being combined. Here Φ_1, Φ_2, etc. can be of different dimensionalities.

4. Experiment Setup

The UBMs utilised in all experiments reported in this paper were trained on the distinct neutral sentences provided for background model training as part of the challenge database [16]. This background dataset consists of speech from 11 subjects with a total of 14 min training data. All the SVM back-ends (in all reported systems) were setup using the WEKA toolkit and utilised the sequential minimal optimization (SMO) algorithm employing a linear kernel with complexity parameter, C, set as either 0.01 or 0.001 [21].

Three different evaluation setups were utilised to evaluate the performances of all the systems outlined in section 3. The primary evaluation setup, herein referred to as the "standard setup", involved training the SVM (task dependent) back-ends using the designated challenge training set and testing the systems on the challenge development set. A second evaluation setup, herein referred to as the "inverted setup", consisted of using the challenge development set to train the SVM (task dependent) back-ends and testing the systems on the designated training set. Finally the third evaluation setup, referred to as the "crossfold setup", did not make use of the development set and evaluated system performance via a 10-fold cross validation on the designated training set. The standard setup results are directly comparable to the development set results reported in [16]. The additional flipped setup and crossfold setup evaluations were performed to detect and avoid any overfitting since all preliminary experiments to determine hyperparameter values were carried out only in the standard setup (testing on the development data). Finally, in order to balance the SVM training data for the readingspanSentence task, the number of training instances across all the three cognitive load classes were upsampled by factors of 3, 3, and 2 respectively using the SMOTE algorithm [22] implementation in WEKA. For the 'inverted setup' the development set was upsampled by the same ratio and for the crossfold setup the number of training samples corresponding to the 3rd class, L3, were cut down to match the number of samples from the other two classes (again only for the readingspanSentence task).

5. Results

5.1. Development Results

The performances of all the individual systems outlined in section 3 were individually evaluated on the challenge development set and the results obtained are reported in Table 1. In addition, a number of combined systems were also evaluated. The performances of the most promising systems as evaluated on the "standard setup" were also evaluated under the "inverted" and "crossfold" setups (refer section 4) in order to determine if the back-end is overfitting. These results are reported in Table 2.

The baseline system results reported in Table 1 are reproduced from the Schuller et. al. [16] since the "standard setup" is directly comparable to the challenge development results. On the other hand, the baseline system results reported in Table 2 ("inverted" and "crossfold" setups) are UARs evaluated using the baseline feature set provided as part of the challenge. For SVM based systems, the SVM complexity parameter (C) was chosen based on performance across all three development setups. The chosen C value for each system is indicated in Table 1 (the same value was used to obtain the results in Table 2).

Table 1: 3-Class unweighted average recall (UAR) on challenge development set – Standard setup along with complexity parameter (C) values for SVM based systems

System	UAR	C
Baseline **(B-sys)** [16]	63.2 %	0.01
Frame-level MFCC+Δ+ΔΔ sub-system	45.2 %	N/A
MFCC+SDC sub-system **(M-sys)**	62.9 %	0.01
SCF+ Δ+ΔΔ sub-system **(S-sys)**	55.9 %	0.01
i-vector sub-system (Total factor = 30)	54.8 %	0.01
JFA sub-system ($N_V = 20, N_U = 5$)	54.8 %	0.01
B-sys + M-sys	67.4 %	0.01
B-sys + S-sys	63.8 %	0.01
M-sys + S-sys	64.7 %	0.01
B-sys + M-sys + S-sys	66.6 %	0.001

Table 2: 3-Class UAR evaluated for "inverted" and "crossfold" setups

System	UAR	
	Inverted	**Crossfold**
Baseline **(B-sys)** [16]	57.2 %	59.8 %
MFCC+SDC sub-system **(M-sys)**	56.3 %	59.2 %
SCF+ Δ+ΔΔ sub-system **(S-sys)**	50.4 %	51.1 %
B-sys + M-sys	61.5 %	64.2 %
M-sys + S-sys	57.5 %	59.4 %
B-sys + M-sys + S-sys	61.6 %	64.8 %

5.2. Test Set Results

The utterance-level combined system including MFCC+SDC supervectors, SCF+Δ+ΔΔ superverctor and the baseline features (B-sys + MS-sys + S-sys) which performed consistently well across all three development test setups was evaluated on the challenge test set and the resultant UAR was determined to be 63.7%. The three individual cognitive load level recall rates were 73.1%, 55.1% and 62.8% respectively.

6. Conclusions

This paper describes our submission to the Interspeech 2014 ComParE cognitive load sub-challenge. The systems were developed with the specific aim of exploring the performance of basic utterance level classification frameworks, based on the cores of most current and past speaker recognition systems. The results show that the utterance level MFCC+SDC performs on par with the baseline system and the SCF+Δ+ΔΔ system complements this MFCC system effectively. The combined system taking into account MFCC+SDC supervectors, SCF+Δ+ΔΔ supervectors and the challenge baseline features outperformed the baseline results on all development and test set results. The comparatively poor performances of the JFA and i-vector based systems may be due to a lack of sufficient training data leading to unreliable estimation of the factor analysis hyperparameters. It is worth noting that the systems evaluated as part this submission were basic frameworks inspired by speaker verification systems and no attempts were made to normalise for speaker and/or other sources of variability (except for a basic attempt with the JFA sub-system).

7. References

[1] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. Van Gerven, "Cognitive Load Measurement as a Means to Advance Cognitive Load Theory," *Educational Psychologist,* vol. 38, pp. 63-71, 2003/03/01 2003.

[2] E. Shriberg, J. Bear, and J. Dowding, "Automatic detection and correction of repairs in human-computer dialog," presented at the Proceedings of the workshop on Speech and Natural Language, Harriman, New York, 1992.

[3] F. G. W. C. Paas and J. J. G. Van Merriënboer, "Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach," *Journal of Educational Psychology,* vol. 86, pp. 122-133, 1994.

[4] P. W. M. Van Gerven, F. Paas, J. J. G. Van Merriënboer, and H. G. Schmidt, "Memory load and the cognitive pupillary response in aging," *Psychophysiology,* vol. 41, pp. 167-174, 2004.

[5] C. Müller, B. Großmann-Hutter, A. Jameson, R. Rummer, and F. Wittig⋆, "Recognizing Time Pressure and Cognitive Load on the Basis of Speech: An Experimental Study," in *User Modeling 2001*. vol. 2109, M. Bauer, P. Gmytrasiewicz, and J. Vassileva, Eds., ed: Springer Berlin Heidelberg, 2001, pp. 24-33.

[6] A. Berthold and A. Jameson, "Interpreting symptoms of cognitive load in speech input," *COURSES AND LECTURES-INTERNATIONAL CENTRE FOR MECHANICAL SCIENCES,* pp. 235-244, 1999.

[7] Y. Bo, C. Fang, N. Ruiz, and E. Ambikairajah, "Speech-based cognitive load monitoring system," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 2041-2044.

[8] P. N. Le, "The use of spectral information in the development of novel techniques for speech-based cognitive load classification," PhD Thesis, University of New South Wales, 2012.

[9] R. Ruiz, E. Absil, B. Harmegnies, C. Legros, and D. Poch, "Time- and spectrum-related variabilities in stressed speech under laboratory and real conditions," *Speech Communication,* vol. 20, pp. 111-129, 11// 1996.

[10] H. Boril, S. Omid Sadjadi, T. Kleinschmidt, and J. H. Hansen, "Analysis and detection of cognitive load and frustration in drivers' speech," *Proceedings of INTERSPEECH 2010,* pp. 502-505, 2010.

[11] T. F. Yap, "Speech production under cognitive load: effects and classification," PhD Thesis, University of New South Wales, 2012.

[12] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication,* vol. 52, pp. 12-40, 1// 2010.

[13] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *Signal Processing Letters, IEEE,* vol. 13, pp. 308-311, 2006.

[14] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 788-798, 2011.

[15] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," 2005.

[16] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, *et al.*, "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load," presented at the Proc of InterSpeech, 2014.

[17] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Proc. of International of Conference on Spoken Language Processing*, 2002, pp. 82-92.

[18] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *Proceedings of ICASSP*, 1998, pp. 617-620.

[19] J. M. K. Kua, T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, "Investigation of spectral centroid magnitude and frequency for speaker recognition," in *Proc. Odyssey Speaker and Language Recognition Workshop*, 2010, pp. 34-39.

[20] M. Kleinschmidt, "Methods for capturing spectro-temporal modulations in automatic speech recognition," *Acustica united with acta acustica,* vol. 88, pp. 416-422, 2002.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.,* vol. 11, pp. 10-18, 2009.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal Of Artificial Intelligence Research,* vol. 16, pp. 321-357, 2002.