High-Level Speech Event Analysis for Cognitive Load Classification

Claude Montacié [1], *Marie-José Caraty*[1]

[1] STIH Laboratory, Paris Sorbonne University, 28 rue Serpente, 75006, Paris, France

Claude.Montacie@paris-sorbonne.fr, Marie-Jose.Caraty@ParisDescartes.fr

Abstract

The Cognitive Load (CL) refers to the load imposed on an individual's cognitive system when performing a given task, and is usually associated with the limitations of the human working memory. Stress, fatigue, lower ability to make decisions and perceptual narrowing are induced by cognitive overload which occurs when too much information has to be processed. As many physiological measures and for a non-intrusive measurement, speech features have been investigated in order to find reliable indicators of CL levels. In this paper, we have investigated high-level speech events automatically detected using the CMU-Sphinx toolkit for speech recognition. Temporal events (speech onset latency, event starting time-codes, pause and phone segments) were extracted from the speech transcriptions (phoneme, word, silent pause, filled pause, breathing). Seven audio feature sets related to the speech events were designed and assessed. Three-class SVM classifiers (Low, Medium and High level) were developed and assessed on the CSLE (Cognitive-Load with Speech and EGG) databases provided for the Interspeech'2014 Cognitive Load Sub-Challenge. These experiments have shown an improvement of 1.5 % on the Test set compared to the official baseline Unweighted Average Recall (UAR).

Index Terms: cognitive load measurement, silent and filler pause, speech onset latency, paralinguistic challenge

1. Introduction

The Cognitive Load (CL) refers to the load imposed on an individual's cognitive system when performing a given task [1]. In Neuropsychology and Psychology, the cognitive architecture summarily consists of a limited and short-term working memory with partly-independent processing units for visual/spatial and auditory/verbal information supervised by a central executive which interacts with a comparatively unlimited long-term memory [2, 3]. The long-term memory stores individual's cognitive schemas of acquired knowledge which are accessible by the working memory to perform a task. From practice or learning, the working memory is able to produce and transfer new cognitive schemas to the long-term memory [4, 5]. Overload can occur when the working memory processes too much information and too fast. A high CL level is known to affect performance effectiveness in achieving a task [6]. The CL measurement is of a great interest to advance CL theory [4, 5] and is used for innovative design in learning [7] and adaptation of mobile conversational interfaces [8].

Three main techniques were used in the research of CL measurement: self-rating, dual-task and physiological techniques. The self-rating technique consists in a subjective rating measure based on the assumption that subjects are able to introspect their CL and report on a scaling rate the cognitive capacity which is allocated for the task [5]; a drawback of this technique is to be intrusive in case of trial series [9]. The dual-task technique is based on the measure of performance in achieving simultaneously a primary and secondary task in order to compare the performance to that of the primary task for a measurement of the interference [5]. In laboratory environment, the secondary task is usually a memory span task using the capacity of the working memory [10]. In such dual-tasks, simultaneously to the primary task, subjects are required to remember a series of items (e.g., letters, figures, words) for a further recall; a variant consists in a mental arithmetic [11].

The Physiological techniques are based on the assumption that changes in CL impact physiological measures such as heart rate variability, eye activity and skin conductance [12]. For its non-intrusive measurement, speech-based features have been investigated in many studies to find reliable CL indicators. Low-level features such as F0, intensity, MFCC and formant were intensively investigated for automatic speech-based CL classification systems [13-16]. Previous studies have shown that high-level features, such as speaking rate, pause characteristics and speech onset latency, can potentially be used in CL level recognition [17, 18].

For the Interspeech'2014 Computational Paralinguistics Cognitive Load Sub-Challenge [19], we paid a particular attention to High-Level Speech Events (HLSE) that should impact classification performance according to related works on CL measurement. In particular, we investigated the response time to stimulus, the speech temporalities (e.g., speech rates) and specific parts of the speech signal (e.g., pause, breathing) which are sensible to cognitive fatigue. The paper is organized as follows. In Section 2, the three speech corpora of the study are described. In Section 3, the Automatic Speech Recognition (ASR) system developed with the CMU-Sphinx toolkit is described. The HLSE transcriptions are analyzed. Features related to HLSE are described and assessed using information gain method. For each corpus, the contribution of the most relevant HLSE feature is analyzed. In Section 4, several CL classifiers using feature sets related to HLSE are designed and assessed on the Development set. The best combination of audio features was investigated. The last section concludes the study.

2. Speech material

The Cognitive Load with Speech and Electroglottograph (CSLE) database [16] was designed for investigations on speakers' CL. CSLE is made of three databases: Stroop Time Pressure, Stroop Dual Task and Reading Span Sentence for which time pressure and dual task technique were used to induce the higher CL level. For the three corpora, speech utterances were labeled into three CL levels: Low (L1), Medium (L2) and High (L3).

For both Stroop Time Pressure and Stroop Dual Task databases, the primary task is the Stroop test [20]: a common test in Psychology which shows a semantic interference phenomenon. The Stroop effect and numerous variants of the test were investigated [21, 22] with a convergence in the results: the response latency to the stimulus increases with the CL level. In the Stroop test, subjects are instructed to pronounce the name of colors displayed on a screen. Two CL

levels are basically induced: (1) low-level of CL, the font color is similar to the color name; this stimulus refers to the congruent test (e.g., when "blue" is displayed in blue font, the subject is required to say "blue"), (2) high-level of CL, the font color is different from the color name; this stimulus refers to the incongruent test (e.g., when "blue" is displayed in red font, the subject is required to say "red"). For the Stroop databases, ten color names (black, blue, brown, gray, green, orange, pink, purple, red and yellow) were chosen. Per trial, the ten different colors (randomly ordered) were all displayed at the same time and repeated twice. At the higher CL level of the Stroop Time Pressure database, a time pressure was added through a color display interval of 0.8 s. At the higher CL level of the Stroop Dual Task database, a secondary tone-counting task was added. Subjects were required to count a specific tone played through headphones for a recall at the end of a trial series. Two tones of respectively 1000 Hz and 2000 Hz were considered, a tone was played in two-second intervals and 200 ms before the next color display, their number of occurrence ranged from four to six. The tone-counting concerned the higher frequency tone (2000 Hz). Table 1 summarizes the experimental conditions for each CL level of the Stroop Time Pressure and Stroop Dual Task databases.

Table 1. *Experimental conditions of the Stroop Time Pressure and Stroop Dual Task databases.*

CL level	CSLE-Stroop with Time Pressure	CSLE-Stroop with Dual Task
L1	Congruent test Color series display	Congruent test 1 s color display interval
L2	Incongruent test Color series display	Incongruent test 1 s color display interval
L3	Incongruent test 0.8 s color display interval	Incongruent test 1 s color display interval Tone-counting

For the CSLE-Reading Span Sentence database [16], the dual task technique was used to induce CL. In this experiment, subjects were required in a series of short sentences to read aloud the sentence displayed and to verify its logic (false or true) while memorizing a series of letters for a recall at the end of the series. For the memory span test, a set of trials was designed to memorize from one to four letters. Each subject read 75 sentences split into (1) five sets of respectively two, three and four sentences, and (2) six sets of five sentences. In each set of trials, a sentence was displayed for reading, then a letter appeared for 800 ms which the subjects were required to memorize. A practice session was provided to the subjects. The CL level labeling was based on the assumption that as the number of letters to be recalled in a set increases, the amount of working memory used will also increase with the CL level. The CL level was associated to the rank *r* of the sentences in the sets of trials for which the number of letters to recall is *r-1*. For any set of trials, Ll was decided for the first sentence (empty recall series), L2 for the second sentence (recall series made of one letter), and L3 for the third, fourth and fifth sentences (recall series made of two, three or four letters). Each subject recorded 21 L1 and L2 utterances and 33 L3 utterances.

26 speakers were recorded for the three databases. Speech data were split into Training, Development and Test sets respectively made of 11 (2 females and 9 males), 7 (2 females and 5 males) and 8 (2 females and 6 males) speakers. For both the Training and Development sets: the Stroop-based databases include nine occurrences of speech utterances per speaker (three utterances per CL level) with respectively an average duration of 16.5 s and 20.8 s for Time Pressure and Dual Task; the Reading Span Sentence database includes respectively 21, 21 and 33 speech utterances per speaker for the three CL levels with an average duration of 4 s. Table 2 gives for the three databases of the Cognitive Load Sub-Challenge the number of utterances of the Training, Development and Test sets.

Table 2. *Number of utterances in the Train(ing), Devel(opment) and Test sets of the databases.*

Database	Train	Devel	Test
Stroop with Time Pressure	99	63	72
Stroop with Dual Task	99	63	72
Reading Span Sentence	825	525	600

3. High-level speech events

Previous studies have shown systematic CL influences on HLSE such as silent pauses, filled pauses, disfluencies and speech onset latencies [17, 23, 24]. HLSE can be extracted from an automatic speech transcription using constraints of the task. We looked for improvement in CL classification using features tied to these events.

3.1. Automatic speech transcription

Speech transcriptions of the databases were obtained using the CMU-Sphinx toolkit [25]. This ASR system may not yet be robust enough for an unconstraint transcription. However, the linguistic characteristics (e.g., lexicon, grammar) of the databases should allow a sufficient robustness of the transcription for the automatic HLSE extraction. Version 0.8 of the Pocketsphinx recognizer library [25] is used to develop the ASR system. The acoustic models were the pre-trained generic US-English acoustic models provided by CMU.

Table 3. *List of the unit transcriptions of the ASR.*

Phonemes
AA AE AH AO AW AY B CH D DH EH ER EY F G HH IH IY JH K L M N NG OW OY P R S SH T TH UH UW V W Y Z ZH
Silent and filler pauses
SIL BREATH NOISE COUGH SMACK UH UM

Table 3 gives the list of units (phonemes, silent and filler pauses) that the acoustic models are able to extract. For the transcription of Reading Span Sentence database, the acoustic models were not re-trained. A <phone | pause> loop search is implemented to get a list of phonemes and pauses with their time-codes. For the transcription of the Stroop databases, the acoustic models are re-trained on the Training set. The phonetic transcription of the color names results from the CMU Pronouncing Dictionary. 14 variants of color name pronunciations were added. A generic word model [26] was introduced and tuned on the Training set. A <color_name | pause | generic_word> loop search was implemented to get a

list of color names, phones, pauses and generic words with their time-codes.

The pause and generic word transcriptions were analyzed. For the Reading Span Sentence database, the total duration of pause segments was 0.78 s in average per utterance. For the Stroop Time Pressure and the Stroop Dual Task databases, the total duration of the segments of pause and generic word was respectively 5.85 s and 9.39 s in average per utterance.

3.2. High-level speech event features

HLSE were analyzed from their statistics on the databases. Five usual statistics of HLSE were computed from the ASR time-codes of pause and phone transcriptions.

- Speaking rate (S): the number of phones per second.
- Articulation rate (A): the number of phones per second, excluding the total duration of silent pauses.
- Pause ratio (P): the total duration of silent and filled pauses divided by the total duration of utterance.
- Filled pause ratio (F): the total duration of filled pauses divided by the total duration of pauses.
- Speech Onset latency (SO): the time interval between the presentation of the stimulus and the first phone (excluding the filled pauses) uttered by the subject.

For the Stroop databases, five other HLSE features were computed from the starting time-codes of the segments labeled as a color name.

- Mean (M): the average of the starting time-codes.
- Standard deviation (SD): the standard deviation of the starting time-codes.
- Kurtosis (K): the Kurtosis measure of the starting time-codes.
- Skewness (SK): the Skewness measure of the starting time-codes.
- Color word number (C): the number of color names uttered by the subject.

For each database, the relevance of all the features is given by the information gain [27] which is computed on the Train set with the following formula:

$$H(class) - H(class/feature) \quad (1)$$

where Shannon entropy H is estimated from a table of contingency and class ={L1, L2, L3}. Features for which the information gain is greater than zero are considered as relevant.

For the Reading Span Sentence database, three features out of five are relevant and selected for the CL classifier. The ranking order of relevance is the following: SO, P and S. Figure 1 shows per CL level the SO histogram computed for each point *t* of the time abscissa as the number of utterances having a SO value in the interval [t, t + 0.1 s]; a spline curve is drawn from the values of the histogram. We remark that the SO distribution is bi-modal: the first mode (around 0.1 s) corresponds to the utterances of L2 and L3 levels, the second mode (around 0.4 s) corresponds to the utterances of L1 level. We notice that the speech onset latency is shorter for high CL level than for low CL level.

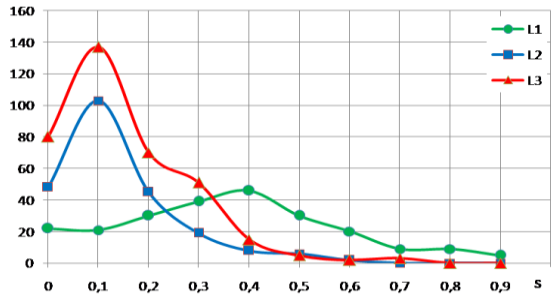

Figure 1: *SO histogram per CL level of the Reading Span Sentence utterances of the Training set.*

For the Stroop Time Pressure database, seven features out of ten are relevant and selected for the CL classifier. The ranking order of relevance is the following: SD, M, S, A, P, C and F. Figure 2 shows per CL level the SD histogram represented as in Figure 1 with a computation interval of 0.4 s. We remark that the SD distribution is tri-modal: the first mode (around 3.2 s) corresponds to the utterances of L1 CL level, the second mode (around 4.5 s) corresponds to the utterances of L3 CL level, and the third mode (around 4.9 s) corresponds to the utterances of L2 CL level. We notice that the SD value is lower for L3 CL level than for L2 CL level. The color display interval only introduced for the L3 CL level seems to achieve the synchronization of the speaker responses to the stimuli.

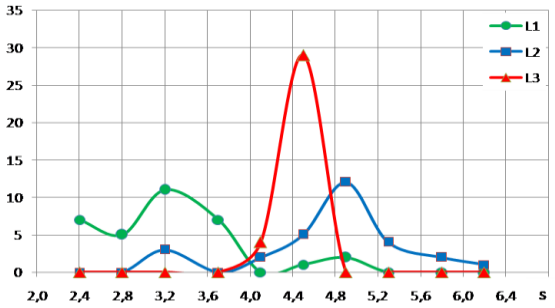

Figure 2: *SD histogram per CL level of the Stroop Time Pressure utterances of the Training set.*

For the Stroop Dual Task database, seven features out of ten are relevant and selected for the CL classifier. The ranking order of relevance is the following: M, F, C, K, S, SK and SO.

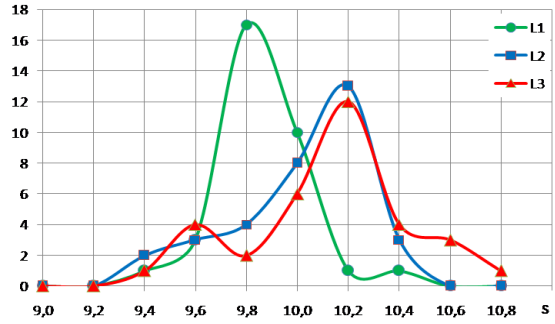

Figure 3: M histogram per CL level of the Stroop Dual Task utterances of the Training set.

Figure 3 shows per CL level the M histogram represented as in Figure 1 with a computation interval of 0.2 s. We remark that the M distribution is bi-modal: the first mode (around 9.8 s) corresponds to the utterances of L1 CL level, the second mode (around 10.2s) corresponds to the utterances of L2 and L3 CL levels. This distribution confirms the results of related work on the Stroop effect: under a high CL level, the response time to the color display increases.

4. Cognitive load classifier

A CL SVM classifier referred to as C1 was developed with the following characteristics: WEKA data mining tool kit [28], Support Vector Machines (SVM) classifier with linear Kernel and Sequential Minimal Optimization, the HLSE feature selection obtained by the information gain method (cf. Section 3.2) referred to as HLSE-IG set for the utterance representation. To account for the imbalanced class distribution of the Reading Span Sentence database, the L1 and L2 categories were up-sampled by a factor of 50%. Table 4 gives the UAR of the C1 classifier on the Development set for the CLSE databases. For the Stroop databases, the UAR is higher in absolute value of 1.4 % compared to the Official Baseline Classifier (OBC). It is noticeable that seven features of the HLSE-IG set gave a better result than the 6,373 features of the official set [19] referred to as IS-2014 set.

Table 4. UAR *on the Development set* of the *CL classifier C1 based on the HLSE-IG set.*

Task	Reading Span Sentence	StroopTime Pressure	Stroop Dual Task
HLSE-IG	{SO, P, S}	{SD, M, S, A, P, C, F}	{M, F, C, K, S, SK, SO}
# features	3	7	7
C1	54.9 %	**76.2 %**	**65.1 %**
OBC	**61.2 %**	74.6 %	63.5 %

These results have shown that the features related to the pauses, pause ratio (P) and filler pause ratio (F), were relevant for CL measurement. Other relevant features could be obtained from the audio analysis of pause segments.

4.1. Audio analysis of pause and phone segments

Two audio feature sets Pa-2010 and Pa-2014 were extracted from the fusion of segments of pause and generic word (for the Stroop databases) of each utterance. Two audio feature sets Ph-2010 and Ph-2014 were extracted from the fusion of the remaining segments (non-pause and non-generic word) of each utterance. The goal is to investigate the CL influence on pause and phone respectively. Pa-2014 and Ph-2014 sets were extracted using the official IS-2014 set. Pa-2010 and Ph-2010 sets were extracted using the IS-2010 set (1,582 features) provided by the organizers of the Interspeech'2010 Paralinguistics Challenge [29]. All of the features were extracted using the open source openSMILE feature extraction tools [30]. Table 5 gives the UAR on the Development set of the two pause-based CL classifiers (Pa-2010 and Pa-2014) and the two phone-based CL classifiers (Ph-2010 and Ph-2014). For the Stroop Dual Task database, the UAR of Pa-2010 CL classifier is higher in absolute value of 4.7 compared to the OBC. It is noticeable that phone-based CL classifiers gave the worst accuracy (except Ph-2014 for Stroop Time Pressure). These results suggest that the CL level has a lower influence on the phone segments than on the pause segments.

Table 5. *UAR on the Development set of the Pause-based and Phone-based CL classifiers.*

Task	Reading Span Sentence	Stroop Time Pressure	Stroop Dual Task
Pa-2010	51.5 %	68.2 %	**68.2 %**
Pa-2014	52.7 %	66.8 %	63.5 %
Ph-2010	46.5 %	63.5 %	55.6 %
Ph-2014	49.4 %	71.4 %	60.3 %
C1	54.9 %	**76.2 %**	65.1 %
OBC	**61.2 %**	74.6 %	63.5 %

The CL classifier referred to as C2 used an optimal combination of the seven audio feature sets {HLSE-IG, Pa-2010, Pa-2014, Ph-2010, Ph-2014, IS-2014} referred as FC2 set. The CL classifier referred to as C2N used a per cluster normalization method using the FC2 set. Clusters are obtained by an unsupervised speaker ID method [31] using the Pa-2010 feature set. Table 6 gives the UAR on the Development set of the C2 and C2N classifiers for the three CLSE databases. The UAR of the C2N classifier was higher in absolute value of 3.1, 6.4, and 15.9 respectively compared to the OBC. These results correspond to an improvement of 4.6 % of the official baseline result (composite result on the three databases) for the Cognitive Load Sub-Challenge.

Table 6. UAR *on the Development set* of the *CL classifiers using optimal combination of feature sets and feature selection.*

Task	Reading Span Sentence	Stroop Time Pressure	Stroop Dual Task
FC2	{HLSE-IG, IS-2014}	{Pa-2010, Ph2014}	{Pa-2010, IS-2014}
C2	61.9 %	80.9 %	73.0 %
C2N	**64.3 %**	**81.0 %**	**79.4 %**
OBC	61.2 %	74.6 %	63.5 %

The best result on the Test set, using the C2 classifier for Reading Span Sentence and C2N classifier for both Stroop tasks, was 63.1 % corresponding to an improvement of 1.5% compared to the OBC. C2 classifier was also used for the Physical Load Sub-Challenge with an UAR improvement of 1.4 % on the Test set.

5. Conclusions

In this paper, we have presented seven CL classifiers which were based on features related to HLSE extracted from an ASR system. Seven feature sets were assessed: one set computed from the transcription time-codes, two sets extracted from the pause segments, two sets from the phone segments and two sets corresponding to the combination of the five previous sets and the official feature set. HLSE features have been shown relevant to CL classification, in particular the speech onset latency, the starting time-code statistics and the audio characteristics of the pause segments. An UAR improvement of 1.5% was obtained on the Test set compared to the result of the Official Baseline Classifier of the Cognitive Load Sub-Challenge.

6. References

[1] van Gog, T. and Paas, F.,"Cognitive Load Measurement", In Encyclopedia of the Sciences of Learning", pp. 599-601. Springer US, 2012.

[2] A Baddeley - Working Memory - Vol. 255 - Issue 5044 - 1992 - pp. 556-559 - Science/AAAS - American Association

[3] Baddeley, A., "Working memory and language: an overview", Science, vol. 255, no. 5044, pp. 556–559, 2003.

[4] Chandler, P. and Sweller, J., "Cognitive load theory and the format of instruction", Cognit. Instr., 8, pp. 293-332, 1991.

[5] Paas, F., Tuovinen, J. E., Tabbers, H. and Van Gerven, P. W., "Cognitive load measurement as a means to advance cognitive load theory", Educational psychologist, 38(1), 63-71, 2003.

[6] Paas, F. and van Merriënboer, J. J. G., " Instructional control of cognitive load in the training of complex cognitive tasks.Educational Psychology Review, 6,551–571, 1994.

[7] Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. Educational psychologist, 38(1), 43-52.

[8] Jameson, A., Kiefer, J., Müller, C., Großmann-Hutter, B., Wittig, F., and Rummer, R., "Assessment of a user's time pressure and cognitive load on the basis of features of speech", In Resource-adaptive cognitive processes, Springer Berlin Heidelberg, pp. 171–204, 2010.

[9] Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload.Human Factors, 33,17–33

[10] A. R. Conway, M. J. Kane, M. F. Bunting, D. Z. Hambrick, O. Wilhelm, and R. W. Engle, "Working memory span tasks: A methodological review and users guide," Psychonomic Bulletin & Review, vol. 12, no. 5, pp. 769–786, 2005.

[11] Lépine, R., Bernardin, S., & Barrouillet, P. (2005). Attention switching and working memory spans. European Journal of Cognitive Psychology, 17, 329-346.

[12] Chen, F., Ruiz, N., Choi, E., Epps J., Khawaja, M. A., Taib, R., Yin, B. and Wang, Y., "Multimodal behavior and interaction as indicators of cognitive load", ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 2, no. 4, p. 22, 2012.

[13] B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah, "Speech-based cognitive load monitoring system," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2008, pp. 2041–2044.

[14] Boril, H., Sadjadi, O., Kleinschmidt, T., Hansen, J.H.L., 2010. Analysis and detection of cognitive load and frustration in drivers' speech. In:Proc. Interspeech, Makuhari, Chiba, Japan, 2010, pp. 502–505.

[15] Le, P. N., Ambikairajah, E., Epps, J., Sethu, V., and Choi, E. H., "Investigation of spectral centroid features for cognitive load classification", Speech Communication, 53(4), pp. 540–551, 2011.

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

[17] Muller, C., Grossmann-Hutter, B., Jameson, A., Jummer, R., Wittig, F., 2001. Recognizing time pressure and cognitive load on the basis of speech: an experimental study. Lecture Notes Comput. Sci., 24–33.

[18] Berthold, A., Jameson, A., 1999. Interpreting symptoms of cognitive load in speech input. In: Proc. Internat. Conf. on User Modeling, 1999, pp.235–244.

[19] Björn Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, Yue Zhang: "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load", Proceedings INTERSPEECH 2014, ISCA, Singapore, Singapore, 2014.

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

[21] MacLeod, C. M., "Half a century of research on the Stroop effect: an integrative review", Psychological bulletin, 109(2), pp. 163–203, 1991.

[22] Linnman, C., Carlbring, P., Åhman, Å., Andersson, H. and Andersson, G., "The Stroop effect on the Internet", Computers in human behavior, 22(3), pp. 448–455, 2006.

[23] Esposito, A., Stejskal, V., Smékal, Z., and Bourbakis, N., "The significance of empty speech pauses: cognitive and algorithmic issues", In Advances in Brain, Vision, and Artificial Intelligence", Springer Berlin Heidelberg, pp. 542–554, 2007.

[24] Khawaja, M. A., Ruiz, N. and Chen, F., "Think before you talk: An empirical study of relationship between speech pauses and cognitive load", In Proceedings of the 20th Australasian Conference on Computer-Human Interaction: Designing for Habitus and Habitat, ACM, pp. 335–338, 2008.

[25] A. Chan, E. Gouva, R. Singh, M. Ravishankar, R. Rosenfeld, Y. Sun, D. Huggins-Daines, M. Seltzer, "The Hieroglyphs: Building Speech Applications Using CMU Sphinx and Related Resources", www.cs.cmu.edu/~archan/share/sphinxDoc.pdf

[26] Bazzi, I., Glass, J.R., "Modeling out-of-vocabulary words for robust speech recognition",Proceedings of ICSLP, Beijing, China, p. 401-404, 2000.

[27] Rauber, T.W., Steiger-Garcao, A.S., (1993). Feature selection of categorical attributes based on contingency table analysis. paper presented at the Portuguese Conference on Pattern Recognition, Porto, Portugal.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann,and I. Witten, "The WEKA data mining software: An update," SIGKDD Explorations, vol. 11, 2009.

[29] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. and Narayanan, S.,(2010). The Interspeech 2010 paralinguistic challenge. paper presented at the Interspeech Conference, Makuhari, Japan, 26-30 September, 2794–2797.

[30] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in opensmile, the Munich open-source multimedia feature extractor," in Proc. of the 21st ACM International Conference on Multimedia, MM 2013, Barcelona, Spain, October 2013, pp. 835–838.

[31] X. Anguera, T. Shinozaki, C. Woofers, and J. Hernando, "Model complexity selection and cross-validation EM training for robust speaker diarization", Proceedings of ICASSP, 4, pp. 273-276, 2007.