Cognitive workload classification with psychophysiological signals for monitoring in safety critical situations

Eydis Huld Magnúsdottir

Doctor of Philosophy

January 2019

School of Science and Engineering

Reykjavík University

Ph.D. Dissertation

Cognitive workload classification with psychophysiological signals for monitoring in safety critical situations

by

Eydis Huld Magnúsdottir

Dissertation submitted to the School of Science and Engineering
at Reykjavík University in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

January 2019

Thesis Committee:

Dr Jon Gudnason, Supervisor
Assistant Professor, Reykjavík University, Iceland

Dr Kamilla Run Johannsdottir, Co-advisor
Associate Professor, Reykjavík University, Iceland

Dr Arnab Majumdar, Co-advisor
Reader, Faculty of Engineering, Department of Civil and Environmental Engineering, Imperial College London, England

Dr Paco Saez, Examiner
Reader in ATM & CNS, Cranfield University, England

The undersigned hereby certify that they recommend to the School of Science and Engineeringat Reykjavík University for acceptance this Dissertation entitled **Cognitive workload classification with psychophysiological signals for monitoring in safety critical situations** submitted by **Eydis Huld Magnúsdottir** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy (Ph.D.) in Electrical Engineering**

25.01 2019

date

Dr Jon Gudnason, Supervisor
Assistant Professor, Reykjavík University, Iceland

Kamilla Johannsdottir

Dr Kamilla Run Johannsdottir, Co-advisor
Associate Professor, Reykjavík University, Iceland

Dr Arnab Majumdar, Co-advisor
Reader, Faculty of Engineering, Department of Civil and Environmental Engineering, Imperial College London, England

Dr Paco Saez, Examiner
Reader in ATM & CNS, Cranfield University, England

The undersigned hereby grants permission to the Reykjavík University Library to reproduce single copies of this Dissertation entitled **Cognitive workload classification with psychophysiological signals for monitoring in safety critical situations** and to lend or sell such copies for private, scholarly or scientific research purposes only. The author reserves all other publication and other rights in association with the copyright in the Dissertation, and except as herein before provided, neither the Dissertation nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatsoever without the author's prior written permission.

25.01 2019

date

Eydis Huld Magnusd

Eydis Huld Magnúsdottir
Doctor of Philosophy

Cognitive workload classification with psychophysiological signals for monitoring in safety critical situations

Eydis Huld Magnúsdottir

January 2019

Abstract

Monitoring cognitive workload has the potential to improve both performance and fidelity of individuals facing safety-critical situations in their working environments. Psychophysiological signals, in particular from speech and the cardiovascular system, are an opportune choice for monitoring individuals providing minimum intrusion and disruption. For reasons perhaps mostly rooted in individual differences, current methods are limited in that moving beyond binary classification has proved challenging. The aim of the present research was to investigate the potential of speech- and cardiovascular signals both separately and in conjunction, for cognitive workload detection, using short-term variability for screen and heart rate based classification schemes. The aim was further to explore alternative methods in order to take into consideration individual differences in the cardiovascular signals that describe the extent of the reactions of the whole cardiovascular system to cognitive workload. For this purpose, new method where a single distance measure describing the hemodynamic reactions of individuals during tasks compared to their own baseline is introduced. A total of 100 university students participated in a study providing extensive information on reactions to cognitive workload that can be used for individual comparisons. The results showed that trinary classification is well achievable with the methods introduced and that the two signals do compliment each other in the cognitive workload classification task. The proposed distance measure showed obvious reactions to cognitive workload during tasks and that these reactions are highly various between individuals.

Flokkun hulægs vinnuálags út frá lifeðlisfræðilegum einkennum fyrir vöktun á vinnuálagi

Eydis Huld Magnúsdottir

janúar 2019

Útdráttur

Með því að fylgjast með huglægu vinnuálagi hjá einstaklingum sem starfa í ábyrgðarmiklum störfum, s.s. flugumferðarstjórn, er unnt að hafa jákvæð áhrif á bæði árangur og líðan þeirra. Lífeðlisfræðileg merki, einkum frá einkennum í tali sem og hjarta- og æðakerfi, eru einstaklega hentugar aðferðir til að fylgjast með einstaklingum og valda þar að auki lágmarks truflun á störfum þeirra. Af ástæðum sem kunna að mestu að eiga rætur að rekja til mismunar á einstaklingum, hefur hins vegar reynst erfitt að flokka þessi merki niður í meira en tvo flokka. Markmiðið með þessari rannsókn var að kanna möguleika á að draga einkenni frá tali annars vegar og hjarta- og æðakerfi hins vegar og nýta þau til að greina þrjá flokka af huglægu vinnuálagi, bæði í sitt hvoru lagi og saman. Markmiðið var einnig að kynna nýja aðferð til að greina og dýpka skilning á mismuni á milli einstaklinga á milli þeirra merkja frá hjarta- og æðakerfi. Í þessu skyni er kynnt ný aðferð þar sem ein mælieining er notuð til að lýsa viðbrögðum hjarta- og æðakerfisins sem einni heild við huglægu vinnuálagi, fyrir hvern og einn einstakling miðað við hvíldarástand hans. Alls tóku 100 háskólanemar þátt í rannsókn sem veitti víðtækar upplýsingar um viðbrögð við huglægu vinnuálagi. Niðurstöðurnar sýndu að vel væri hægt að flokka í þrjá flokka með þeim aðferðum sem kynntar eru og að þessi tvö lífeðlisfræðilegu merki bættu hvort annað upp við flokkun á huglægu vinnuálagi. Augljós einstaklingsbundin viðbrögð við huglægu vinnuálagi komu í ljós milli hvíldarástands og verkefna sem kröfðust mikils huglægs vinnuálags.

I dedicate this work to my husband Lárus and my children Ástrós Lilja and Ólafur Magnús, I could not have done this without you.

Acknowledgements

This work was funded by the Icelandic aviation authority ISAVIA and Icelandair. The experiment design and data collection phase of this work was funded by The Icelandic Centre for Research (RANNIS) under the project *Monitoring cognitive workload in ATC using speech analysis*, Grant No 130749051.

.

Preface

This dissertation is original work by the author, Eydis Huld Magnusdottir.

The work introduced in Study I was initiated by Manuela Meier [1], but presented at the cognitive infocommunications (CogInfoCom), 2016 7th IEEE international conference and extended for publication in the journal periodica polytechnica electrical engineering and computer science by Magnusdottir *et al.* [2].

List of publications

M. Meier, M. Borsky, E. H. Magnusdottir, K. R. Johannsdottir, and J. Gudnason, “Vocal tract and voice source features for monitoring cognitive workload,” in Cognitive Infocommunications (CogInfoCom), 2016 7th IEEE International Conference on, 2016, pp. 000097–000102.

E. H. Magnusdottir, M. Borsky, M. Meier, K. Johannsdottir, and J. Gudnason, “Monitoring Cognitive Workload Using Vocal Tract and Voice Source Features,” Periodica Polytechnica Electrical Engineering and Computer Science, May 2017.

E. H. Magnusdottir, K. R. Johannsdottir, C. Bean, B. Olafsson, and J. Gudnason, “Cognitive workload classification using cardiovascular measures and dynamic features,” in 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Debrecen, 2017, pp. 000351–000356.

K. R. Johannsdottir, E. H. Magnusdottir, S. Sigurjonsdóttir, and J. Gudnason, “Cardiovascular monitoring of cognitive workload: Exploring the role of individuals’ working memory capacity.,” Biological psychology, 2017.

E. H. Magnusdottir, K. R. Johannsdottir A. Majumdar, and J. Gudnason, “Cognitive workload classification with heartbeat synchronized cardiovascular and voice features,” Pending review.

Contents

List of Figures

List of Tables

List of abbreviations

ANS	Autonomic nervous system
ATCO	Air traffic control operators
BDM	Bhattacharyya distance measure
CE	Classification entity
CT	Confusion table
ECG	Electrocardiogram
EEG	Electroencephalography
EM	Expectation-maximization algorithm
GMM	Gaussian mixture model
HR	Heart rate
KARMA	Kalman-based autoregressive moving average model
MCR	Misclassification rate
MTR	Mistrust rate
NASA-TLX	NASA task load index
PSD	Power spectral density
PNS	Parasympathetic nervous system
PSF	Performance shaping factors
RF	Random forest
SA	Sinoatrial node
SD	Standard deviation
SE	Standard error of the sample mean
SNS	Sympathetic nervous system
SVM	Support vector machines

Part I

Overview

Chapter 1

Introduction

Monitoring cognitive workload has the potential to improve both the performance and fidelity of human decision-making. Cognitive workload is essentially a theoretical concept adhering to the mental effort needed to complete a task contingent on the individuals' competence to perform it. Aviation is an ideal example of an industry where a real-time cognitive workload monitoring model would be beneficial to assist human operators in their crucial role of ensuring safe and efficient air travel.

One of the challenges with monitoring cognitive workload is to locate the optimum measure that reflect the individual's reactions to cognitive workload reliably in real-time. The signals traditionally measured can be grouped in three; (1) performance signals, (2) self assessment questionnaires and (3) psychophysiological responses of the human body, e.g. eye movement and heart rate.

The use of signals from the cardiovascular system such as heart rate and blood pressure are well known for their role in the monitoring of individual long term health [3]. The autonomous nervous system (ANS) regulates the cardiac cycle, both during normal state and during instinctive fight-or-flight reactions to danger [4]. This regulatory system is complex, with contradictory but compensatory functions making interpretation of signals from the cardiovascular system with respect to cognitive workload problematic. Cardiovascular signals might seem intuitively ideal for the purposes of cognitive workload monitoring and in many ways they are. Researchers have reported binary classification (e.g. low/high or neutral/high) with excellent results but have, however, not been able to move beyond that to more granular levels (see e.g. [5]). The main reason for this seems to be that traditional methods have failed to account for the different reactions to cognitive workload particular to each individual.

Although not traditionally associated with cognitive workload monitoring, characteristics in features extracted from speech signals have been used to detect up to three distinct cognitive workload levels [6]. By applying the methods used for traditional speech processing tasks such as speech transcription, speaker identification and language detection to the task of cognitive workload detection the assumption is that speech signals also contain information regarding cognitive workload levels.

The aim of this research is to develop a model using speech- and cardiovascular signals for classification of distinct cognitive workload levels to enable monitoring of workload in safety critical situations. The research questions of this thesis are the following;

- Can cognitive workload be detected from speech- or cardiovascular signals and if so can these signals be incorporated to move beyond binary distinction?

- Do these two psychophysiological signals compliment each other to improve cognitive workload classification?
- Are there methods that can better help us to understand the complex reactions of the cardiovascular system and thus to better monitor cognitive workload?

For this purpose a trinary classification scheme is introduced through three sets of studies with an additional study aimed at identifying profiles representing groups of individuals and tasks according to their cognitive workload reactions. The first study is dedicated to exploring methods of cognitive workload classification with two groups of feature sets derived from speech signals. Shifting the focus towards the cardiovascular signals in Study II, a method for short-term variability calculations and heartbeat classification scheme are introduced. Concluding these sets of studies, speech- and cardiovascular signals are combined in Study III to explore their complimentary function in cognitive workload classification. A method for measuring the extent of cognitive workload reactions through one distance measure from baseline (normal state) to task periods is introduced in Study IV. Groups of individuals are then identified according to their hemodynamic reactions through profile value feature sets retrieved from the distance measure for each task. For validation of the proposed methods for cognitive workload classification presented in the studies, speech- and cardiovascular signals from 100 participants were collected during an extensive experimental session. First, however, some of the fundamental concepts and methods as well as the state-of-the-art research in the relevant fields is introduced.

1.1 Cognitive workload

In the highly technologically driven world of today, more and more working environments demand high levels of cognitive workload. Especially in safety-critical fields such as aviation, surgery and landing crane operation, where the monitoring of the human operators' cognitive workload becomes extremely important [7].

The concept of cognitive workload refers to the relationship between task demands on the one hand and the cognitive capacity required to cope with these demands on the other [8]–[11]. When task difficulty surpasses capacity, performance breaks down, making the individual more vulnerable for producing errors [12].

Cognitive workload studies focus on detecting cognitive over-/under load for the purpose of avoiding their detrimental effect on performance and attention. Overload occurs when the task load exceeds the working memory resources of the individual [7]. Cognitive under load, on the other hand, refers to situations where the cognitive demands of the task become so low that the individual becomes susceptible to distractions and experiences reduced attention [13]. The effect of cognitive under load may become more important with the increasing automation of tasks and the human tasks become more monotonous and homogeneous [14].

The opaque nature of cognitive workload makes constructing measurable units for it a challenge. This is caused by the huge amount of counteractive factors affecting human performance, described by e.g. the concept of performance shaping factors (PSF). PSF is a common concept in the study of human error in aviation and may also be applied to the general study of cognitive workload. The PSF in aviation are according to Boring *et al.* [15]: (1) Training and experience, (2) procedures, (3) ergonomics and human machine interaction, (4) time available, (5) complexity, (6) workload and stress, (7) environment, (8) fitness for work and (9) work processes. This example from aviation reflects the challenge of evaluating cognitive workload. The PSF become, in this case, 9 influencing dimensions with internally complex conjunctions and dependencies. Furthermore, these dependencies are highly dynamic with the constant risk of safety-critical situations occurring with short notice.

1.2 Measuring cognitive workload

Cognitive workload in itself is a theoretical construct that cannot be measured directly. Rather, it is measured by (1) the use of subjective ratings scales, (2) performance measures associated with primary and secondary tasks or (3) by psychophysiological measures. Most of these methods have not been applied to the real-world operational environment [7], [16], but rather have been used in experimental environments.

1.2.1 Subjective workload assessment

Subjective rating information is traditionally gathered in the form of questionnaires administered at intervals throughout the task period. Their aim is to capture the multidimensional nature of cognitive workload through the subjects direct estimate of the workload experienced at a given time [17]. The subjective workload assessment technique [17] and the NASA task load index (NASA-TLX) [18] are both examples of methods widely used.

The more widely used of the two, the NASA-TLX rating scale was designed to incorporate task-, behavioral- and subjective- related experiences. It consists of six scales designed to capture cognitive workload levels that can be applied within and between task evaluations in a wide variety of fields [18]. The six scales are (1) mental-, (2) physical- and (3) temporal demand along with scaling off (4) performance, (5) effort and (6) frustration level. Each component is reflected through closed ended questions accompanied by a rating of e.g. high/low or good/bad response options. NASA-TLX is relatively easy to interpret and can be adjusted by weights to reflect the needs of the perspective rater [18].

The form of execution of subjective rating scales through questionnaires makes their applicability in the field for continuous monitoring of workload hard to implement [7], [9]. Furthermore, the sensitivity of subjective ratings as a general workload scale is debatable [19]. In general, rating scales rely on the individual's judgment of their functional state but it is clear from the literature that individuals have very limited access to their cognitive states (e.g. [20], [21]).

1.2.2 Performance based measures

Performance measures are typically grouped into primary- or secondary tasks according to whether the measure depends on the actual working tasks or an additional side task explicitly introduced to evaluate performance [7], [22].

Primary task measures focus on measuring the extent of achievement of goals set for either operator or system performance [7]. The goals set depend on the task and situation and might need to be defined and adjusted to each situation. The variety of tasks and situations, especially in the field of aviation, makes the implementation of primary task measures a complicated one. The operator might have several primary tasks and changes in priorities throughout different stages of the task can also complicate matters [7].

Secondary tasks performance measures involve the solving of a separate quantifiable task in addition to the main task and the effect of time sharing is measured from the perspective of either one of the tasks [22]. Either the performance of the secondary task is used as criteria for cognitive workload state called subsidiary-task paradigm or alternatively load-task paradigm measuring the performance of the primary task [7], [22]. Secondary task performance evaluations have traditionally been applied in the empirical environment as e.g. an aid to evaluate the capacities of the human operator [7].

Although in many cases easily applied, primary task measurements are not suited for complex dynamic working environments where multiple contradictory situations can occur at short notice. In the case of secondary tasks, developing tasks that do not affect the performance of the primary task is a challenge [22]. Therefore a more direct approach, measuring the psychophysiological responses of e.g. air traffic controllers (ATCO) might prove to be more effective for aviation cognitive workload monitoring.

Researchers have found that primary performance measures [23] as well as [24] performance on secondary tasks successfully distinguished between levels of task difficulties. However, it should be kept in mind that the connection between workload and performance is a complex one and although intuitively, increased task load should lead to increased workload this is not necessarily the case. Capacity of the individual operator varies in terms of level of expertise, as well as, fatigue, stress and emotional states. The operator's strategy can also adapt to the demands of the task and may become better with experience [25].

1.2.3 Measures from psychophysiological signals

The notion of detecting cognitive workload from psychophysiological signals is based on the premise that cognitive workload is manifested in physiological reactions. Psychophysiological signals introduce an exciting alternative to both subjective- and performance based measures as they offer objective measurement in real-time and for most parts, depending on the measure, in a relatively non-intrusive manner. The most commonly researched measures for cognitive workload detection are electrical brain activity [26], [27] and cardiovascular reactivity [28]–[31].

The brains role of receiving, processing and administering sensory information makes electrical signals ideal for reporting the brains engagement to cognitive workload [32]. Equipment such as electroencephalogram (EEG) are used for recording electrical activity of the brain by employing intrusive multiple sensors positioned on the scalp. Functional near infrared spectroscopy is a non-intrusive alternative for detecting brain activity, but is a relatively new method still not suited for commercial use (see e.g. [33]).

The cardiovascular system has been proven time and again to show significant reactions to outside stimuli and the long term negative affects of continuous stress is also well known [3]. Furthermore, it is known that an increased task demand will affect the cardiovascular system in the short term [34], [35]. Traditionally, cardiovascular signals are recorded with an electrocardiogram (ECG) using electrodes placed on the torso.

In aviation, the ideal and least intrusive psychophysiological measure would be through the verbal communications between ATCO and pilots. Whilst not applicable in all environments e.g. surgical theater with all present wearing mouth masks, it is ideal in situations where speech can be captured relatively unobtrusively and communications can be monitored in real-time without interruptions.

In general, psychophysiological measures have quite a few advantages that can be beneficial for real-time cognitive workload monitoring. They provide objective measures free from the individual's perception, they are multidimensional, implicit, continuous and have been proven to be responsive to stimuli [36]. Cumbersome and expensive equipment for signal collection, dependency on massive data acquisition and computationally heavy interpretation methods are among the disadvantages [36]. However, with technological enhancements such as wearable devices and ever increasing computational capacity of computer systems, these measures are now becoming a more feasible option [37].

Both speech- and cardiovascular signals seem especially promising for cognitive workload monitoring. Speech signals can be collected through vocal communications and cardiovascular signals through inconspicuous wearable devices ensuring the minimum intrusiveness of the monitoring in real-life situations. Separately, the efficacy of these signals in detecting cognitive workload has been proven [6], [38] and can in conjunction potentially compliment each other.

Not only is the understanding of the methods traditionally applied to extracting features from each of these intended signals needed for the development of a cognitive workload monitoring model. In addition, at least some knowledge of the human anatomy behind the signals is essential for a successful interpretation. The next two sections are dedicated to describing the two signals of importance for this work with respect to cognitive workload classification modeling. We will go through the main anatomic functions behind the signals and the methods traditionally used to extract meaningful features from the two.

1.3 Speech signal processing

Monitoring cognitive workload through speech communications between e.g. pilots and ATCO is a fundamental concept in this research. If applicable, recording communications can provide a non-intrusive way to monitor cognitive workload in employees facing safety critical circumstances. Many factors are revealed through the act of speaking. The most obvious factor is the linguistic content of the speech but factors such as identity, gender, health and emotion can also be deduced from speech signals. In addition, speech signals have in recent years also been proven to reveal cognitive workload [6], which is the focus of this work.

1.3.1 Speech production

As illustrated in Fig. 1.1, the human voice production system contains three elements; (1) the lungs, (2) the vocal folds within the larynx and (3) the articulators. The vocal folds contain elastic ligaments stretched between the rigid cartridges of the larynx and intrinsic laryngeal muscles varies the tension in the vocal folds, much like loosening and tightening a guitar string [4]. Air from the lungs passing through the larynx vibrates the vocal folds producing sound pressure waves traveling through to the articulators,

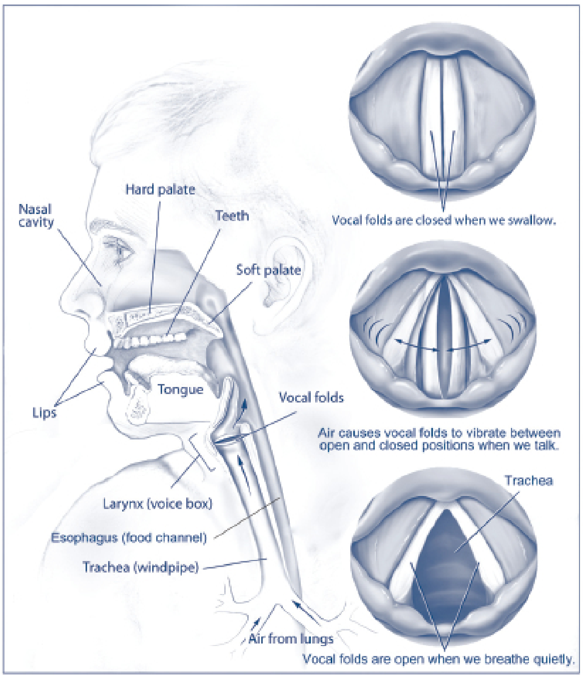

Nasal cavity
Hard palate
Teeth
Soft palate
Lips
Tongue
Vocal folds
Larynx (voice box)
Esophagus (food channel)
Trachea (windpipe)
Air from lungs
Vocal folds are closed when we swallow.
Air causes vocal folds to vibrate between open and closed positions when we talk.
Trachea
Vocal folds are open when we breathe quietly.

Figure 1.1: A diagram of the human speech production system. On the left the three elements of the system from the air flow from the lungs through the larynx and to the articulators above the larynx are illustrated with a cross section of a human head and neck. On the right three functional positions of the vocal folds illustrate the role of the larynx in swallowing, sound production and normal breathing with a top down view of the larynx. *Figure copy right ©JohannO14/Wikimedia Commons/Creative Commons Attribution-Share Alike 4.0 International license.*

producing the final sounds (phonation). The articulators are the part of the vocal production system above the larynx; the pharynx, nose and mouth that produce recognizable and individualized human speech [4]. The pitch of the sound pressure wave is controlled through the degree of tension in the vocal folds and the pressure of air flow controls the loudness. Vowels are produced by relaxing and contracting the muscles in the pharynx and enunciation is aided by muscles in the face, lips and tongue [4]. These sound pressure waves travel to the speaker's immediate environment and are received by listeners or perhaps a microphone. The interpretation of the sound pressure waves recorded is the task of speech signal processing.

1.3.2 Speech signal processing

Audio signals are a continuous stream of sound pressure waves emanating from their origin to their receiver. The frequency of the human voice typically ranges from $300-3000$ Hz. However, the human receiver only partially distinguishes the information within each signal rendering the capture of the entire signal for reproduction redundant. Thus samples of the continuous signals, at a sufficiently fast rate, become an accurate enough representation of the original continuous signal [39]. In speech processing, the continuous sound pressure wave is sampled to produce a discrete signal $s[n]$ at a sampling frequency which can vary, typically from $8-48$ kHz depending on application.

In frame-based processing of speech the discrete speech signal $s[n]$ is then converted into a data matrix where each row contains an M-dimensional vector $\mathbf{s}_t = [s[t], s[t+1], \ldots, s[t+M-1]]^T$, where t is the index of the start of the row vector in the discrete speech signal and M is the size of the time frame which typically corresponds to $25-30$ ms. The time index t can be chosen with an increment of 1 (sliding window) but the typical hop-size used is 10 ms. The resulting data matrix therefore contains row vectors of 30 ms frames which can be analyzed separately. This choice of window size is a trade-off between frequency resolution which is achieved by having the frame size long and being able to assume that the frame is a stationary signal which you can do if the frame size is short enough.

There are many approaches to extracting features from the speech signal that have the potential to detect the cognitive workload level of speakers. One approach would be to treat the audio signal as an unknown entity and simply modeling the signal directly without preprocessing and hope for interesting results. These methods depend on huge amount of data without the underlying assumptions and reasons for the results being known. The popular pattern recognition methods of neural networks, and Gaussian mixture models (see Section 1.5) are a few of these *brute force* methods [40]. The goal of this research is not only to achieve good classification results but also to gain insight into their underlying function and therefore other methods are employed in this work.

1.3.3 Vocal tract features

Vocal tract features is a term used for the first three formant frequencies representing resonant regions, depicting the slowly varying changes in the shape of the vocal tract due to articulation. The vocal tract features have been proven to vary depending on mood and have been used successfully for cognitive workload classification [41]. Traditional methods for formant feature extraction focus on estimating the spectral components of the voice by transforming the sound pressure wave from time- to frequency domain. An example of feature tracks can be seen in the spectrogram (top) portion of Fig. 1.2 as darker shades of bands or regions that differ in density throughout the utterance.

Spectral analysis of the vocal tract signals entails the analysis of preprocessed windows of the signals with the goal of describing their distribution over frequency of the power contained therein [42]. The results are information containing the amplitudes of the resulting spectrum for each window. Figure 1.2 depicts an example of a speech wave and its corresponding spectrogram. This segment is taken from the validation database used in this research where a female speaker utters the word 'red' in Icelandic (rauður). In the lower part of the figure the differences between vowels and consonants can be seen. For example, the alveolar |r| can bee seen from 0.05-0.13 seconds and blending into the diphthong |Ö||í| (au) between 0.14-0.23 seconds. These differences can also be seen in the corresponding spectrogram (top) where the period for the alveolar differs from that of the diphthong. The alveolar has higher energy in lower frequencies but the diphthong has clearer formant tracks. Collectively, the information in all the windows is then used to characterize the whole speech signal.

In speech processing, parameter estimation techniques are based on the assumption that there is an underlying stationary stochastic process that produces the speech waveform that can be described using a small number of parameters [43]. The task becomes to estimate these parameters so that this stochastic process is described.

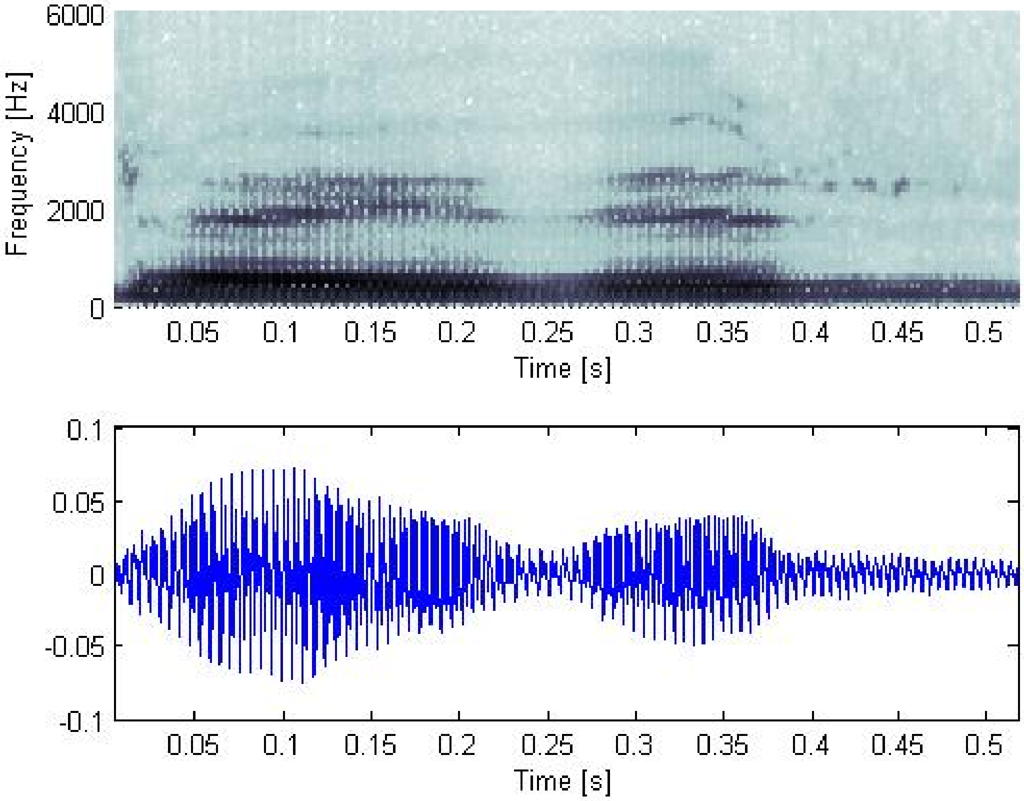


Figure 1.2: An example of a speech wave (bottom) and its corresponding spectrogram (top) with the frequency formant tracks (dark horizontal regions). The example is taken from a segment where a female speaker utters the word rauður (red) in Icelandic.

By applying the framing and pre-processing steps described earlier, this assumption becomes true for the spectrum of the frame $\tilde{\mathbf{s}}_t$.

An example of a modeling method based on stationary stochastic processes is autoregressive processing (AR). Adding the second polynomial to the algorithm, moving average (MA), a complete set of the auto-regressive moving average model is employed in this work. The two polynomials are typically applied to time series, either separately or in conjunction depending on application, to provide a description or sometimes prediction of a weakly stationary process. The AR provides a prediction on the next successive point in the series through a regression of it's past values and MA provides a model of the error terms as a linear estimation of the same values and some various points in the past. The ARMA method is quite common in speech processing and has been used for example in speech coding in voice communications and coding in text to speech synthesis [43], [44]. By applying a time-frequency transform (e.g z-transform or fast-Fourier transform) to the sound pressure wave the ARMA algorithm is applied to estimate formant track features.

The specific algorithm used here (KARMA) employs a Kalman-filter [45], also known as linear quadratic estimation, for point estimates and uncertainty inference for more accurate formant tracking. The KARMA algorithm has been proven to exhibit lower root-mean-square error compared to tested methods as introduced by Mehta *et al.* in 2012 [46] (see Section 3.2.1 for description).

1.3.4 Voice source features

Voice source features is a term used for a group of features that use glottal flow estimates to extract the fast varying changes that transpire within the glottal source. Glottal source analysis focuses mainly on extracting elements such as (1) the duration of each laryngeal pulse (open/closed) periods, (2) the instant of glottal closure,(3) the spectral structure of each glottal pulse and (4) the pulse shape. To achieve this methods such as glottal inverse filtering are employed, where an estimate of the vocal tract filter is found and its inverse applied to the speech signal [47].

Voice source features essentially are interpreting the sounds produced with airflow from the vocal tract thought the glottis to the articulators [47], [48]. In the human speech production system, the glottis consists of (1) a pair of folds of mucus membrane, (2) the vocal folds in the larynx and (3) the opening between the vocal folds [4]. As opposed to vocal tract features, features that are extracted from the voice source are not under conscious control of the speaker and therefore can reveal subconscious mental states and emotions.

1.4 Cardiovascular signal processing

Compared to speech signals, employing cardiovascular signals for cognitive workload detection has a long tradition. Even though the two are in their essence just signals that can be interpreted mathematically, the systems they are retrieved from are quite different and the methods used have to reflect that. In the case of the cardiovascular system it is the interpretation of the autonomous nervous systems complex regulation of bodily functions with respect to cognitive workload that is of importance. Despite this there are opportunities in applying methods cross fields, provided the understanding of the underlying physical processes are accounted for.

1.4.1 Cardiac regulation

Auto-rhythmic cells (pacemakers) regulate the heart rate (HR) but do not determine it. Their role is to set the rhythm for the contraction of the heart, forming the cardiac conduction system [4]. The sinoatrial (SA) node initiates cardiac action potentials 90 to 100 times per minute. Although the initiation of the heartbeat originates in the SA node, the actual HR is controlled by the autonomic nervous system (ANS) [49]. ANS has two major divisions: (1) the sympathetic nervous system (SNS) and (2) the parasympathetic nervous system (PNS). As opposed to the somatic nervous system, operating under conscious control, the ANS usually functions without conscious control [4].

The SNS and PNS function in opposition of each other, not in an antagonistic but complementary way [4]. The SNS is responsible for increasing the HR above the intrinsic rate controlled by the SA node and is typically activated in response to increased outside stimuli e.g. in fight-or-flight situations. The reaction time to an onset of sympathetic stimulation is up to 5 seconds and the system reaches a steady level within 20-30 seconds [50]. The PNS, however, slows down the intrinsic rate of the SA node to a normal or level state. The time of response after onset of PNS stimulation on the SA node is typically within one heartbeat, depending on its phase, and it affects only one or two heartbeats after its onset [4].

Cardiac regulation is managed thorough the medulla where sensory information about limb positions, blood chemistry, state of the heart and the limbic system with information from higher brain centers is collected [49]. Based on this information the medulla sends messages that cause shifts to the relative balance between PNS and SNS adjusting the HR [49].

The reactions of this complex system using information from an array of bodily functions is hard to interpret solely through the signals, in a top down manner, without deeper understanding of the underlying counteracting functionality (see further discussion in Section 2.2).

1.4.2 Cardiovascular variability

Heart rate is the term used for the measurement of oscillation between two consecutive heartbeats and the term for variation between instantaneous HR is heart rate variability (HRV) [3]. Calculating the variability of HR gives insight into the velocity and acceleration of the measure thereby including the dynamic impact of larger time segments than one heartbeat [3]. HR has been proven to be sensitive to cognitive demands, time restrictions, uncertainty, attention and correlated with arousal [36]. HRV on the other hand, has been employed as a measure of mental workload and for positive and negative valence of an experience [36].

Theoretically, the variability of any cardiovascular measure (e.g blood pressure or stroke volume) can be calculated in the same manner as HR; and indeed that is the method applied in this research. Here we look at the cardiovascular signals from a more comprehensive perspective incorporating ten measures describing a more complete hemodynamic profile (see Section 3.2.2). The measures are sampled for heartbeats as opposed to continuously, as with signals recorded with ECG, perhaps resulting in loss of precision. The theory is, however, that the breadth of this hemodynamic profile is more beneficial to cognitive workload, as PNS and SNS onset happen in measures of heartbeats. For the same reason, to capture the variability of the hemodynamic profile sorter segments should be used to reflect variability than traditionally suggested (2-5 minutes) [3]. To test these theories a method known from speech signal processing is used, the delta method, where variability from 2 adjacent heartbeats and acceleration from 2 adjacent variability measures.

1.5 Feature analysis with pattern recognition

Pattern recognition methods are commonly employed to interpret the feature sets extracted from the speech- and cardiovascular signals. For the prediction of classification, we can suppose a feature set matrix $\mathbf{X}_N$ with a total number of N observations and information regarding cognitive states ω_N, such that there exists ω_n to each corresponding x_n feature value vector. A training set can be drawn from $\mathbf{X}_N$ containing $I = N - L$ number of observation with $L \neq N$, denoted $\tilde{\mathbf{X}}_I = [x_1, ..., x_I]^T$ and a corresponding subset of cognitive state observations from ω_N, $\tilde{\omega}_I = [\omega_1, ..., \omega_I]^T$. The goal is to make predictions using a prediction method $y(\tilde{\mathbf{X}}_I, \tilde{\omega}_I)$ with the observations $\tilde{\omega}_I$ and the training set $\tilde{\mathbf{X}}_I$ based on the test set values $\hat{\mathbf{X}}_L$. Ideally the classification result for an unknown value $\hat{x}_l$ produces an accurate prediction $y(\hat{x}_l) = \hat{t}_l + \epsilon$ (with $\epsilon = 0$) regarding the corresponding cognitive state $\hat{\omega}_l$.

Another prediction action commonly taken in pattern recognition is clustering where groups or clusters are identified without any previous knowledge regarding the corresponding states $\hat{\omega}_l$. These methods either identify groups through the distance between data points (e.g k-means) or by determining the distribution of the data (e.g. GMM) [51].

1.5.1 Classification procedures

A classification entity (CE) is the portion of data behind one classification result and is generally represented by one vector of features x_n and its corresponding labels ω_n [52]. During the feature extraction stage this entity might have been reduced from a much larger set of data and even contain different concatenated feature sets. In the case of Study I, CE corresponds to the segment of one screen of speech features. For Study II and III the CE are two; (1) heartbeat and (2) screen containing several heartbeats. The CE in Study IV consist however, of the collection of screens that make up one task (e.g. reading or Stroop levels see description in Section 3.1).

The concept of closed- and open-form refers to whether a definite unique solution can be reached with the chosen classification method [51]. Consequently this raises the question of whether to omit instances (open-form) where definite classification results can not be reached or to come up with other solutions. These solutions might be to use another classification method to solve the outstanding feature sets [51] or, as in this work, using the measurement of the certainty of the classification (soft score) as a decider, denoted $y_k(x_n)$ where k is the class index.

Overfitting is a concept used where the training set uses more adjustable parameters than are optimal for the feature set, i.e. a quadratic target function with four parameters where a linear function with three parameters is sufficient [52]. The result of a overfitting model is that it does not become a good predictor of the test set $\hat{\mathbf{X}}_L$ and the training phase is likely to demand more information for each item than the optimal solution needs [51].

The poor predictive performance of unseen data due to overfitting can become less of a problem if the amount of data is plentiful. Then the solution can simply become to train many models or train with many values and then compare these performances with a separate independent validation feature set. However, set division methods are greatly governed by the amount of data available to divide the feature set.

For the cross-validation method, e.g. if S denotes the number of groups then $(S-1)$ number of groups are used to train the model and the one group left out is used to asses the performance of the model. This procedure is then repeated S times for all possible choices of left-out-test-groups. Thereby the cross-validation method allows for $(S-1)/S$ of the available data to be used for training S times. In the case of very small feature sets a leave-one-out strategy can be applied where the number of groups becomes $S = N$, where N equals the number of data points [51].

In classification the test set data $\hat{\mathbf{X}}_L$ is used to evaluate the performance of the model and a uniform method of presenting the correctness of the pattern recognition algorithm is important. A confusion table (CT) depicts the accuracy of each class in the diagonal line of a matrix and the inaccuracy in the off diagonal parts [53]. The columns of the table represent the result of the classification from the unseen test set value $\hat{x}_i$ and the rows the actual observations $\hat{\omega}_l$ corresponding to $\hat{x}_l$.

The misclassification rate (MCR) or its inverse, correct classification rate (CR) is the most common accuracy score used for presenting results. It depicts the percent of incorrect classification of either all classes or within individual classes. If MCR is to be interpreted as the probability of making an error, there needs to be an assessment of the accuracy of the MCR itself. While standard deviation (SD) is an estimation of the variability of the MCR with repeated experiments the standard error of the sample mean (SE) is an estimate of how much the sample mean will vary from the SD of the sampling distribution [54]. The SE is a calculation of the population size and its SD and displays lower values as number of experiments increases. So, to indicate the uncertainty around the estimate of the mean measurement of the MCR, we quote the SE [54]. Here, the uncertainty of the estimate, or SE, is reported together with the average MCR results for all participants.

A less used accuracy score is the mistrust rate (MTR), indicating the chance of an error if an unknown entity is classified as a certain class.

Having introduced the procedures used for classification the next logical step would be to introduce some of the methods commonly employed for data analysis. The prediction methods chosen as well as the procedures used depend on the nature of the feature sets and the purpose of the analysis being conducted.

1.5.2 Pattern recognition prediction methods

There are quite many well established computer algorithms that have been applied to the task of cognitive workload feature evaluation [55]–[57]. Five classification and clustering methods are introduced which are all used in this work except for neural-networks.

1.5.2.1 Support vector machines

Support vector machines (SVM) is an example of an effective prediction used for classification analysis. Essentially, SVM projects the data into higher dimensional spaces with the aim of making it linearly separable [33]. This concept is based on the theoretic phenomenon that data should be linearly separable if mapped into sufficiently high dimensions [52]. The aim is to estimate a function $f : \mathcal{R}^M \rightarrow \{-1, +1\}$ using M dimensional training set $\tilde{\mathbf{X}}_I$ and class labels $\tilde{\omega}_I$. With the objective of the function $y(\tilde{\mathbf{X}}_I, \tilde{\omega}_I)$ being able to correctly classify an unknown test set $\hat{\mathbf{X}}_L$. Consequently, a class of hyperplanes are defined using the inner product $\langle \cdot \rangle$ as,

$$\langle g, \tilde{\mathbf{X}}_I \rangle + b = 0 \tag{1.1}$$

$$g, \tilde{\mathbf{X}}_I \in \mathcal{R}^n, b \in \mathcal{R} \tag{1.2}$$

corresponding to a decision function:

$$w(x) = sign(\langle g, \tilde{\mathbf{X}}_I \rangle + b) \tag{1.3}$$

becomes the basic process of SVM. The optimal hyperplane can be found uniquely by solving a constraint quadratic optimization problem [58]. The optimal solution is the one that has the largest margin between the two boundaries, separating the two classes and the data points closest to the boundaries are called support vectors. These support vectors contain all the relevant information about the decision boundary. This represents an important property of SVM, that the determination of the model parameters corresponds to a convex optimization problem, so any local solution is also a global optimum [51], i.e. the support vectors alone are able to find the same solution as the whole feature set.

A kernel function can be applied to extend the linear classification of SVM to a non-linear one. When applying the kernel function the inputs are mapped into high-dimensional feature spaces but instead of computing the full list of features for each data point, only the kernel functions are computed [58]. I.e. using information of the relevant position of the data points (inner product) rather than storing information on their exact positions within some reference frame [58]. Thus solving the obstacle caused by one of the drawbacks of SVM, the heavy computational cost of working directly in high-dimensional spaces.

1.5.2.2 Random forest

In pattern recognition, random forest (RF) algorithms are used for solving classification problems. The algorithm comprises of an ensemble of tree, each of which votes for a class based on its decision criteria [59]. The final decision is then the aggregate over the ensemble which is usually based on the most popular decision.

RF is a specific type of bootstrap aggregated (bagged) trees while other examples of ensemble learning methods using more than one hypothesis, are boosted trees and rotation forest. These methods use different schemes to build multiple hypothesis, common element of all these procedures, that select random vectors to govern the growth of each tree in the ensemble can be described. For the rth tree a random vector $\mathbf{w}_r$ that determines the tree's decision criteria is generated with the same distribution but independent of past random vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{r-1}$. The training set $\tilde{\mathbf{X}}_I$ and the random vector $\mathbf{w}_r$ are used to grow a tree (hypothesis) $y(\tilde{\mathbf{X}}_I, \mathbf{w}_r)$. These trees vote for the most popular class when a large number of trees has been generated. The strong law of large numbers can be used to prove that RF trees always converge, eliminating overfitting (see Breimann [59] for details).

Most pattern recognition algorithms use one hypothesis for the best approximation of the assumed underlying function of each vector of features in the training set. The main strength of ensemble algorithms is that they output more than one hypothesis, thus partially overcoming bias and variance problems that other learning algorithms suffer [60]. However, outcome variance can be caused by there being many similar hypothesis that give the same accuracy but the chosen hypothesis might not be the best predictor of the test set $\hat{\mathbf{X}}_L$, especially with large training sets $\tilde{\mathbf{X}}_I$. Equally if the best hypothesis can not be found, computational methods to find the local minima can get stuck and fail to find the best hypothesis. Variance can hence be reduced with ensemble methods using all these hypothesis to vote or by adding weights to several different local minima. Bias occurs when no hypothesis can be found that is a good approximation of the function in the hypothesis space. To account for this a weighted assembly of hypothesis might be used to find a better estimation of the function, thus reducing the bias [60].

1.5.2.3 Neural networks

The concept of neural networks originates from efforts to find mathematical representations of information processing in biological systems [51]. When applying neural networks the number of basis functions are fixed in advance but the parameter values become adaptive during training. A neuron is the most elementary computational unit in the network. It sums a number of weighted input and passes the results through an activation function which is generally non-linear [43].

Neural networks usually work quite well in speech recognition classification, learning complex and non-linear training data. However, the main drawback to the method is its tendency to overfitting [43] when not enough data is available, resulting in poor predictive performance of the model. This is the case in this work and therefore it was decided not to use this method.

1.5.2.4 Gaussian mixture models

Mixture models are a group of probabilistic methods that can be used to identify sub-groups from a set of unobserved data in multidimensional space. In the case of the Gaussian mixture model (GMM) the underlying assumption is that the data can be represented with the Gaussian distribution. The mixture is performed by superposition (weights) of a mixture of Gaussian distributions by taking linear combinations, weighted by a coefficient [51]. GMM has been applied both for classification and clustering predictions differing only in the approaches taken regarding input data and validation.

Given a Gaussian mixture distribution,

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \mathbf{\Sigma}_k) \tag{1.4}$$

where π_k represents the k^{th} latent variable or the mixing coefficient. The parameters for mean μ_k, covariances $\mathbf{\Sigma}_k$ and mixing coefficients π_k are typically determined by maximizing likelihood with the expectation maximization (EM) algorithm [51]. The EM algorithm is an iterative process used for maximum likelihood estimation of the parameters that best represent the statistical values of a random variable based on the observations of incomplete data [43]. After initialization of the mean μ_k, covariances $\mathbf{\Sigma}_k$ and mixing coefficients π_k a measures for their responsibilities (E-step) is found with,

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \mathbf{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \mathbf{\Sigma}_j)} \tag{1.5}$$

The responsibilities represent the value of the posterior probabilities associated with data point x_n. In the M-step the parameters μ_k, $\mathbf{\Sigma}_k$ and π_k are re-estimated. The convergence of the algorithm is then checked after each EM step with

$$\ln p(\mathbf{X}|\mu, \mathbf{\Sigma}, \pi) = \sum_{n=1}^{N} \ln \Big\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \mathbf{\Sigma}_k) \Big\}. \tag{1.6}$$

If the convergence criterion is not satisfied the EM step is repeated.

1.5.2.5 K-means clustering

Unlike the GMM algorithm, k-means is a non-probabilistic method for cluster analysis in multidimensional space. The objective of the k-means algorithm is to find groups, or clusters, of feature sets whose inter-point distances are small compared to the distances to points outside the cluster [51]. If a feature set is defined as $[x_n, \ldots, x_N]$ with N

observations in D dimensions and the total number of clusters K. The objective function becomes

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2 \tag{1.7}$$

where the goal is to find values for r_{nk} (phase 1.) and μ_k (phase 2.) that minimize J. The two phases are conducted with a iterative procedure repeated until convergence or i.e. repeated until no changes are made to the assignment of clusters or some maximum predefined number of iterations is reached [51].

1.6 Chapter summary

Throughout this chapter the fundamental concepts needed to build a comprehensive cognitive workload monitoring model based on blood-pressure data and speech recordings have been introduced. Some of the methods implemented in this thesis have already been proven to perform well in similar situations and some have been applied in different fields of research with good results. The main challenge in this work is to combine input of multiple disciplines making in-depth specialization of each contributing discipline challenging.

In general, there are a few challenges that need to be addressed before a functional cognitive workload monitoring model can become a reality. The most opportune measures need to be chosen and features extracted from them that can be best used to detect cognitive workload. Not only is the challenge methodological, there is also limited understanding of the human bodies complex reactions to outside cognitive stimuli. There are psychophysiological signals that can be detected, but the profiles of individuals reactions to cognitive workload are not clear. A definition of what is too high and what is too low cognitive workload has not been reported and more granularity in classification is needed.

Chapter 2

Research overview

The extensive research reported in the field of cognitive workload motoring using psychophysiological signals is reviewed here. The research efforts laying the groundwork for the work introduced in this thesis is of special interest. The aim is to explain the motivation behind the methods used and some of the choices made during the process. Some of the methods have been previously used for solving similar tasks but quite few were sought from outside fields thereby contributing to cognitive workload monitoring research.

2.1 Speech signals for cognitive workload classification

The notion of employing speech signals to detect cognitive workload levels is relatively new compared to other psychophysiological signals e.g. cardiovascular signals. Research efforts in this field have focused on applying feature extraction methods traditionally used for other speech processing tasks such as speech recognition (see [42], [61] for method descriptions) and speaker recognition e.g. [43].

Research groups related to the aviation industry identified early on the opportunities of using speech through vocal communications as a way to monitor cognitive workload [16]. In an effort funded by NASA [16], strong indicators of characteristics in speech that change during stress were identified using commercial of the shelf speech- and speaker recognition systems, available at the time. Their findings that these systems failed to address the wide speaker variability associated with speech under stress reinforces the theory that cognitive workload can be detected from speech [16].

There are a few speech signal production resources that have the potential to indicate cognitive workload from the speech signal. Pitch, duration, intensity, voice source- and vocal tract spectrum using linear and non-linear feature extraction methods are all examples of possible sources [16]. A significant research effort towards cognitive workload detection from speech signals was performed by researchers at the national ICT Australia (NICTA) and from the University of New South Wales in Sydney [6], [41], [62]–[65]. Their collection of research was based on two validation databases with 15 and 14 participants undertaking reading and comprehension tasks as well as the Stroop task [6] (see Section 3.1.1 for detailed description). All the methods represented trinary classification schemes using GMM unsupervised learning and speaker normalization. For lack of data a uniform background model was trained using reading data to supplement the leave-one-out cross-validation dataset division method used. The research approach of the collection of publications by the NICTA group was to present the results for each speech feature extraction method separately or in pairs or groups. The classification results from their six published papers on the subject are gathered in Table 2.1.

Table 2.1: Research results from the NICTA group on cognitive workload classification from speech signals. The correct classification results CR[%] for each of the speech feature extraction methods tested for open and/or closed set division are listed when applicable.

Author	**Feature Extraction**	**Closed**	**Open**
Yin *et al.* 2008	MFCC and prosodic	77.5	58.5
Yap *et al.* 2010	Glottal features	84.4	NA
Yap *et al.* 2011 a	Formants 1-3	67.7	
Yap *et al.* 2011 b	Voice source features & Formants 1-3	62.7	
Le *et al.* 2010	Voice source & vocal tract filters	80.8	54.5
Le *et al.* 2011	Spectral centroid frequency & -amplitude	88.5	NA

This set of publications covers many of the feature extraction methods used in the speech processing field, giving a good overview of the potential use of speech signals for cognitive workload monitoring.

Surprisingly, high accuracy results were reported in the research for trinary classification with speech signals. The set division method seems to be the reason for this, a *closed set* training process with the speakers used in the training set were also included in the test set [6]. This aberration can, however, easily be overlooked and the results from the *open set* classification used for comparisons. Both open and closed set results are listed in Table 2.1 where applicable and in one column in unclear cases (Yap *et al.* 2011 a and Yap *et al.* 2011 b).

Many of the same pattern recognition methods as described in Section 1.5 have been employed for classification of features extracted from speech signals. Methods such as neural networks, GMM and SVM seem to be most popular with the addition of e.g. Hansen [16] employing Bayesian hypothesis testing and distance measures classifications. The winning entry of the Interspeech 2014 Cognitive Load Challenge (ComParE) [66] used a method called i-vector classification with a combined feature set of fused speech streams, prosody and phone-rate [67].

As an alternative to the traditional delta coefficients (Section 3.2.3), commonly applied in speech processing, to capture temporal variability Williamson *et al.* introduced the temporal correlation structure method [68] in 2012. The method was designed to classify EEG signals for seizure prediction but was extended to speech processing in conjunction with the KARMA method in [46]. This method was employed for cognitive workload classification in Study I. In Study I, features extracted from two speech production sources; the vocal tract and voice source were tested for indications of cognitive workload reactions, on their own and fused at various stages of the process. A trinary classification scheme was tested with three distinct classifiers SVM, RF and minimum distance measures with screen based classification. The classification was performed for individual participants to reduce the impact of a major confound in the field caused by individual differences (see discussion in Section 2.4) and thereby setting the focus on testing the proposed methods.

In Study I trinary classification of low, medium and high cognitive workload states is successfully applied on individual basis. This establishes that granularity can be achieved in classification from speech signals, beyond the state-of-the-art. Furthermore, that for advances in the field to become a reality, the focus needs to be on accounting for or including differences between individuals methods for cognitive workload monitoring.

2.2 Cognitive workload classification with cardiovascular signals

Researchers have proven time and again that cardiovascular signals can be successfully used to detect cognitive workload levels [31], [69]. Jorna [70], for example, found that cardiovascular measures were well suited to index different mental states and dynamic responses to variations in workload. Although widely used successfully for cognitive workload detection, researchers in the field have reported conflicting results regarding the efficacy of both HR and HRV as cognitive workload detection parameters [23], [31], [32]. In a study by Wilson [32], both HR and HRV proved to be consistent for the same pilot measured on two different days, however, neither HR nor HRV distinguished well between different segments of flight (22 in total). Indicating that as cognitive workload detection parameters, HR and HRV are good for detecting whether the individual is performing a task, but not for distinguishing the finer workload changes. Further supporting this, Vogt *et al.* [31] found HR to significantly detect increased number of aircrafts in two different simulations (en-route and tower) and increased number of conflict in en-route simulation. HR however, did not distinguish higher load of vertical traffic or pilot error in en-route simulation nor predictable or unpredictable conflict in tower simulation. Brookings *et al.* [23] found that HR did not distinguish differences in task load (number of aircrafts handled) in a simulated air traffic control task and Kaber *et al.* [71] also found that HRV did not consistently distinguish between high and low traffic load.

In addition to there being confounds between the evident HR and HRV responses to cognitive workload, moving beyond binary classification (low/high or neutral/high) has proven to be difficult [72]. Even though average correct classification results from e.g. 89% [29] to 98% [5] have been reported establishing that cardiovascular measures detect well both task onset and offset [72]. Different approaches, however, need to be explored for distinguishing between adjacent levels of increasing or decreasing levels.

One approach could be to investigate more diverse signals than HR and HRV from the cardiovascular system such as attempting to isolate the cardiac regulation responses from the ANS to cognitive workload [73]. Backs [74] found that features extracted reflecting reactions of the ANS can be used to augment HR as a cognitive workload measure. Other researchers argue that methods need to focus on short-segment analysis to account for the compensatory reactions of the ANS. To this affect, reducing the window size to 30 s., Stuvier *et al.* found that workload manipulation showed strong effects from variability measures of HR and blood pressure [28].

Some studies have focused on identifying groups of individuals according to characteristic associated with cognitive workload. Thereby seeking to explain the confounding challenges evident in the state-of-the-art regarding individual differences. Researching working memory capacity [75] or performance [76], [77] from cardiovascular signals with regard to cognitive workload are examples of these. Another approach also commonly used is to employ pattern recognition methods for individual grouping tasks as well as cognitive workload levels classification. The primary methods used in cognitive workload evaluation research with cardiovascular signals are neural networks and SVM [56]. Comparing the performance of the two methods, Elkin *et al.* [56] found that although neural networks achieved higher accuracy during binary classification, the SVM algorithm achieved greater speed and efficiency.

In Study II the focus was shifted from speech signals to methods for cognitive workload classification from cardiovascular signals. As an alternative to the traditional temporal variability methods such as HRV, a method from speech processing called delta coefficients was applied to the cardiovascular signals. With this method the velocity for 2 adjacent heartbeats is calculated and then in turn the second order acceleration of the two adjacent velocity measures. By focusing on each heartbeat as units with 10 cardiovascular parameters reflecting a wide hemodynamic profile the danger of loss of information due to large windows sizes is addressed with this method. Again trinary classification, now with the two best performing classifiers (SVM and RF) from Study I, was used for analyzing of the cardiovascular signals. In addition to screens as classification entities the results from heartbeat classification is also reported where the classifiers attempted to identify the cognitive workload level from information contained for only in one heartbeat.

The fact that the classification results from Study II (20.44% MCR) outperform the ones from Study I (33.5% MCR) is perhaps not surprising as research in the field of cardiovascular signals is years ahead of the speech signals with respect to cognitive workload. The results in Study II however move beyond the binary results previously reported and in compliance with the state-of-the-art results.

2.3 Combined psychophysiological signals

Researchers studying the state-of-the-art in physiological measures of air crew mental workload realized as early as 1979 that employing more than one physiological measurement could be beneficial for cognitive workload level detection [78]. Methods to assess the performance of two or more psychophysiological signals compared to self assessment or secondary task has been the subject of quite many studies (e.g. [55], [79]). In some instances comparing multiple measures including performance, subjective (TLX), and psychophysiological measures (EEG, eye blink, heart rate, respiration and saccade) [23]. Researchers have also focused on comparing the performance of two or more psychophysiological signals [9], [33], [78] for detecting cognitive workload levels. Quite a few studies have also focused on comparing multiple psychophysiological signals for classifying cognitive workload level for their applicability in aviation [5], [32], [80].

Researchers have also reported methods where two or more psychophysiological signals have been combined to form one feature set. The most prominent have focused on cardiovascular signals combined with electrical brain activity signals either as a pair [26], [81] or grouped with e.g. galvanic skin response signals [82] or oculomotor signals [83], [84]. No attempts, to our knowledge, have been reported investigating the supplemental possibility of cardiovascular- and speech signals in the cognitive workload detection task.

The most straight forward method of combining feature sets from different sources would be simply to concatenate them. There are however a few issues that need to be addressed before a concatenated feature set can be successfully used as a pattern recognition classification input. The sampling rate of the signals might be various so there might not be equal amount of values at the same point in time. Their alignment in time has to be ensured during data recordings as well as their correspondence after individual feature extraction is concluded. Ryu and Myung [83] used factor analysis and multiple regression analysis to create weight coefficients to represent the signals.

Combining features from speech- and cardiovascular signals and aligning them to mutual time segments for the purpose of cognitive workload classification, is the subject of Study III. As described in Sections 3.1.6 and 3.1.7 the sampling rate of the two signals do not match and therefore a time alignment method is introduced for segmenting speech features equidistant before and after heartbeats. Statistical evaluations of the speech segments were then combined with the information from the cardiovascular signals to form a concatenated heartbeat based feature set. Continuing with some of the best performing methods from Study I and Study II the third study focuses on investigating whether speech- and cardiovascular signals could be used to compliment each other for greater accuracy.

2.4 Confounding issues in cognitive workload research

One of the major confound in research attempting cognitive workload detection, is rooted in the fact that human beings do not react in the same manner to cognitive workload. The methods used can therefore not be used to interpret cognitive workload reactions in a general manner without approaching or accounting for this confound whatever the measure [85]–[87]. Researchers have approached the problem of individuality from different perspectives. Some have included methods that reduce this impact into their processes e.g. classification on individual basis [2], [88], thereby focusing on the performance of their proposed methods. Effort has been applied to seeking to explain the underlying reasons for this by identifying characteristics of groups of individuals according to certain criteria. Indeed, as far back as 1968, Hecker *et al.* [89] found that measured voice parameters differed between individuals, where workload was detected in those parameters for some individuals and not others. Grassman *et al.* [85], found that young male pilot applicants who scored high on cognitive avoidant coping style showed less increase in HR in response to workload compared to others.

Researchers have also focused on the inherent reactions of the cardiovascular system to cognitive workload. The intricate reactions of the cardiovascular system to onset of messages from the ANS [74], [90] and automatic baroreflex activation models [73] are among these. The problem is that the interaction between the parasympathetic and sympathetic system is both complex in a counteractive but compensatory manner [73], [91]. In the short run (within 1-5 heartbeats), different parts of the sympathetic system will cause increased cardiovascular reactivity in response to increased demand. In the long-term, blood pressure is regulated through the baroreceptor system that activates the PNS, lowering the cardiovascular reactivity as the blood pressure rises [73]. The hemodynamic profile-compensation deficit HP-CD model was introduced by Gregg *et al.* in 2002 [92] for locating individual differences from cardiovascular signals. The method relates blood pressure regulation as a compensatory relationship with respect to the individual's baseline measures. This method has been used to successfully link individual difference in cardiovascular reactivity to personality trait characteristics such as type D, neuroticism, and depression (e.g. [93]). Johannsdottir *et al.* [75] also showed by using the model that individual working memory capacity may play a critical role in determining how individuals react to changes in cognitive workload.

The research presented in Study IV introduces a novel method that describes through one measure the magnitude of the cardiovascular systems reactions to tasks from the individuals normal state. By assuming that two segments of multidimensional cardiovascular signals are normally distributed a measure describing the distance between the two are calculated with a method called the Bhattacharyya distance. Apparent reactions to cognitive workload tasks are presented in Study IV from this distance measure when sliding windows of segments are compared to the individuals baseline period. Furthermore, common characteristics based on profile features describing four tasks are then used as criteria for identifying groups with similar profiles.

2.5 Chapter summary

In general, it can be stated that changes in cognitive workload is reflected through both the speech production and cardiovascular systems. However, researchers struggle to move beyond detecting high and low cognitive workload to a more reliable measure that can detect small workload changes. The problem seems to be both theoretical and methodological in nature. A deeper understanding of individual reactions to cognitive workload as well as how they are reflected in the psychophysiological signals is needed. In subsequent chapters the contribution of this work to cognitive workload detection research is introduced through four Studies. First however, the validation database design and the methods used in more than one study are described in detail.

Chapter 3

Methods

The research performed in this thesis is based on a vast data collection project funded by the Icelandic centre for research fund (Rannís). The aim of the experiments was to provide recordings of speech- and cardiovascular signals to validate the research questions of whether *cognitive workload can be detected from speech- and cardiovascular signals beyond binary distinction* and whether *their combined feature streams could improve the classification results.* In addition the need for methods that can better help us to understand the complex reactions of the cardiovascular system became apparent early on. Thus the search for methods that can better explain these reactions to cognitive workload also became a part of the work.

3.1 Experiments and data

100 participants, mostly students at Reykjavik University, visited the laboratory for a session of tasks lasting within one hour. The result is a comprehensive multidimensional database which is only partly addressed in this thesis. For completeness, however, the whole experiment setup with all tasks will be included in the description with emphasis on the relevant sections.

3.1.1 Cognitive workload levels - the Stroop task

The main task of the experiment was the well established cognitive word/color task introduced by Stroop in 1935 [94]. During the Stroop test the task is to utter aloud a set of color names, such as 'blue' and 'red' from a computer screen in either congruent color (the word 'red' is shown in red), or in-congruent color (the word 'red' is shown in, for example, blue). For the level of difficulty to be such as to induce cognitive workload reactions in participants the degree of congruency of words vs color matches are adjusted between Stroop levels. Time constraints, in addition to the in-congruency levels, then comprised the third difficulty level with one color appearing on the screen at a time. The specific setup of the three levels with various levels of congruency, in-congruency and time limits are as follows;

Level 1 Seven congruent sets of screens with all 36 color names appearing on the computer screen at the same time.

Level 2 Six in-congruent sets of screens with the alternating two levels of 0.3 and 0.7 in-congruency with all 36 color names appearing on the computer screen at the same time.

Level 3 Eight sets of screens with one word appearing at a time, in the timed intervals of 0.75 s. and 0.65 s. Here the same in-congruency set-up was applied as in level 2 and the same amount (36) of color names per screen as in level 1 and 2.

Each screen j contained a set of 36 words appearing in a 6x6 matrix on a computer screen. The participants task was to utter the color of the words aloud, not to read the color names. In the experimental setup the Icelandic color names for blue, green, brown, red and pink were chosen. The average time to complete a single screen of the congruent task was 23 seconds., increasing to 30 seconds. for the in-congruent task, and 29 and 25 seconds. for the time-limited tasks. The number of screens in each cognitive workload level were chosen in advance so that the participant would spend approximately the same amount of time on each cognitive workload level.

3.1.2 Working memory capacity task - OSPAN

The participants went through the operation span (OSPAN) [95] task which is designed to give score for an individual's working memory based on their performance. During the task the participants are asked to memorize words in between performing another task of calculating simple mathematical equations. The OSPAN task was implemented with alternating screens appearing in front of the participant. The first depicting a simple equation with the answer being either correct or incorrect. After the participant establishes whether the answer is correct or incorrect, a single word appeared for a short period of time. When the predefined number of alternating equation/word sets was complete the participant was asked to recall the words in the correct order. The number of equation/word sequences where initially two and gradually increased to five at a time.

3.1.3 Reading task

A reading task was included in the experiment for the purpose of having a task that the participants were relatively comfortable and familiar with. The text was chosen from a textbook taught in Icelandic schools intended for 10th (14-15 years of age) grade teaching. An intermediate level of difficulty was chosen such that every participant would be challenged but still comfortable with the task and be able to conduct it in a fluent manner. All participants roughly managed to finish the reading of the text within 2 minutes.

3.1.4 Self-assessment questionnaire

A self-assessment questionnaire was administered four times (see Fig. 3.1) throughout the session. The questionnaire used was based on five of the six NASA-TLX [18] as described in Section 1.2.1, translated into Icelandic. The assessment of physical demand of the tasks proved redundant as the experiment was designed to reduce physical movement and was therefore excluded. The five remaining dimensions of the questionnaire were therefore administered with a rating scale of 1-10.

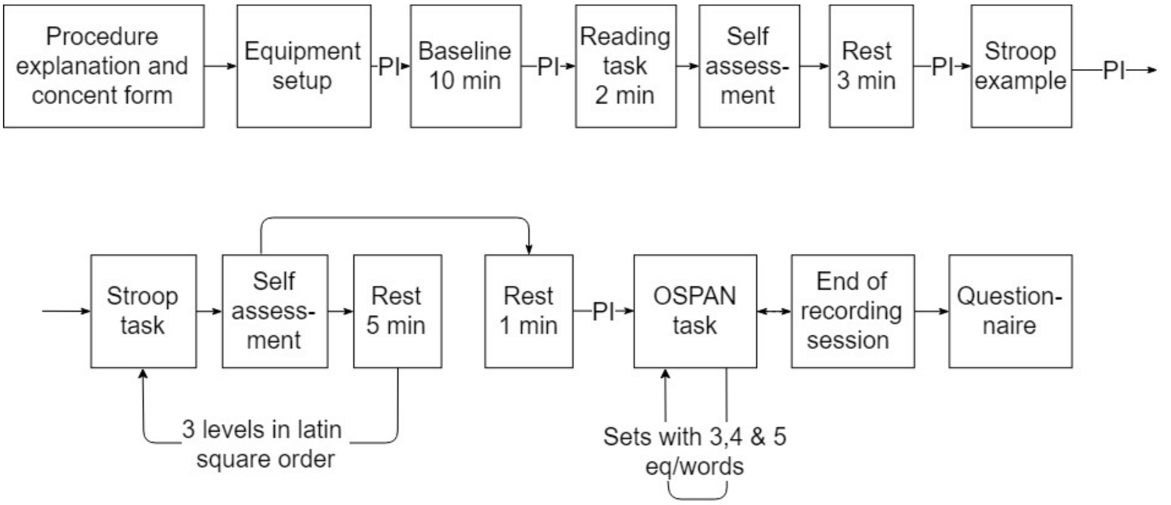


Figure 3.1: The sequence of tasks and resting periods for the whole duration of one experiment session. Progress instructions (PI) signify instructions given to the participants in between tasks which they were asked to be read out loud.

3.1.5 Experiment session configuration

The session started with a short instructions from the researcher and consent form signature. The Finometer Pro, cardiovascular signals measuring device, was then connected and calibrated for a recording through the entire experimental session. All speech was recorded during the experiment by using both a head-mounted and a table-top microphone. The recordings were stored in linear pulse code modulation files with 48 kHz sampling frequency and 16 bits per sample. The participants were asked to read all instructions aloud, throughout the session, to ensure comprehension of their content and for additional speech recordings with the session conducted entirely in Icelandic. They were also asked to keep their movements to a minimum restricting them to verbal responses to reduce interference of physical activity on the recordings.

A detailed step-by-step flowchart of the experiment session is depicted in Fig. 3.1. The first step was a recording of a cardiovascular signal baseline, followed by the reading task and the associated self-assessment questionnaire. The Stroop tasks were introduced with an example of one screen of color names appearing in the color black. The order of Stroop levels was alternated using the Latin square technique where he participants could be presented with combinations of cognitive levels in orders such as $\{L1, L2, L3\}$, $\{L2, L1, L3\}$ or $\{L3, L2, L1\}$ etc. As depicted in Fig. 3.1 each Stroop level containing the Stroop task, self assessment questionnaire and resting period, for the three levels resulting in total number of $J_p = 21$ screens. Fig. 3.1 also depicts the strategic positions of resting periods, designed to ensure that the participant had sufficient time to recover between tasks in order to eliminate the influence of the preceding task. A set of instructions ensured the seamless progress of the experiment and the same level of comprehension between participants. When these tasks where completed the recordings of cardiovascular and audio data was over. The participant was asked to answer a questionnaire designed to give information on general health, sleeping habits, anxiety and depression.

3.1.6 Speech measuring devices and recordings

The speech signals were recorded through two microphones, one mounted on the participants head and another directly in front of them. The sampling frequency was $f_s = 44.1$ kHz with a typical length of speech segment that was analyzed 35 seconds, which means that the discrete signal vector is of the length (44100x35) samples. The audio recordings were stored in separate audio files for each change in session.

3.1.7 Cardiovascular measuring device and recordings

The Finometer Pro from Finapres was used to record the cardiovascular responses of the participants during the experimental sessions [96], [97]. This device is based on the volume-clamp method developed by Penáz, patented in 1967, and the physiocal criteria of Wesseling [97]. The signals are obtained using a finger cuff and an upper arm cuff for calibration of the reconstructed blood pressure. The Finometer was set to start immediately after set-up and calibration was complete at the beginning of the baseline and ended after the last OSPAN task was completed, in one continuous recording. Although, the Finometer PRO takes continuous measurements of the cardiovascular signals the supporting computer program, BeatScope [98], produces measurements and their derivations for each heartbeat n, making them consistently uneven in time (see detailed description in Section 3.2.2). The sensitivity of the finger cuff measuring device to movement or interruption, in the detection of blood flow to the finger, results in there being periods where the cardiovascular signal recording is interrupted. These periods are accounted for by applying the average of the four surrounding measurements to replace the one missing.

3.1.8 Database setup

Before any data processing was initiated a complete database setup was designed using the database administration tool MySql. The schema for the database is illustrated in Fig. 3.2. This preprocessing of the data has ensured streamlined data analysis for researchers and made it more accessible to new researchers.

As inevitable in a data recording of this caliber, some of the recordings were faulty due to various reasons. For example during one experiment session, interruption in the Finometer recordings resulted in a gap in the cardiovascular signal recordings. Although affecting the Studies containing the cardiovascular data, this did not influence the audio recordings and therefore this particular contribution could be included in the speech signals analysis study. These instances turned out to be surprisingly few keeping the number of participants to $P = 98$ in Study I, $P = 97$ in Study II and $P = 96$ in Study III.

3.2 Data processing methods

In this section the data processing methods used in more than one study will be introduced and methods specific to a single study are introduced in their corresponding chapters. These are feature extraction, temporal extraction, concatenation and classification methods that have proven to give good classification results either from research with similar objectives or from lessons learned throughout this process.

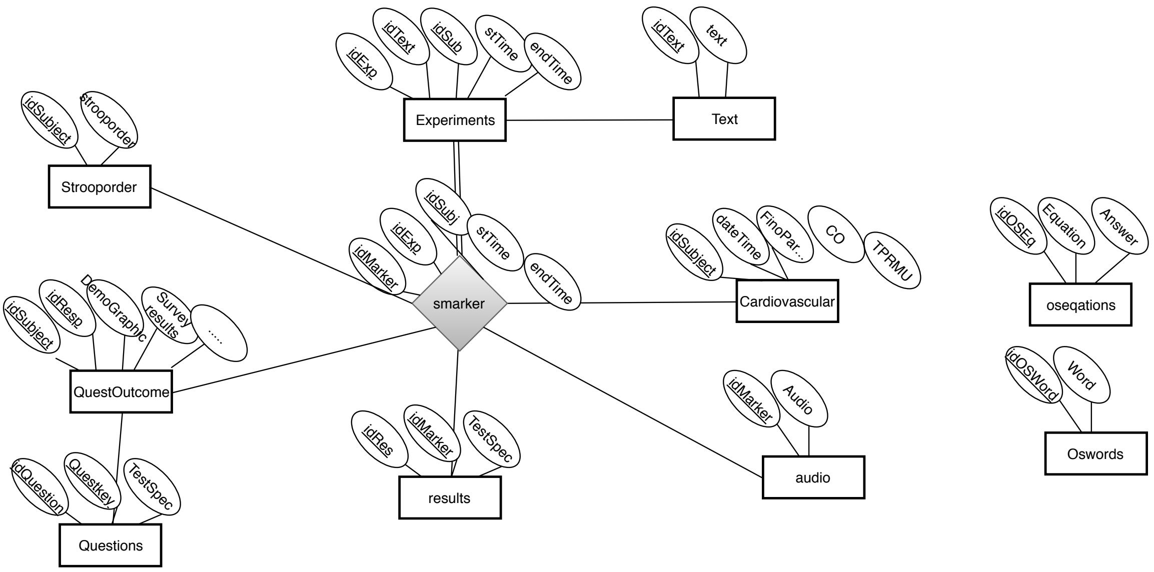


Figure 3.2: Schema depicting the setup and connections of the tables of the experimental database.

3.2.1 Vocal tract feature extraction

There are numerous means of extracting useful features from speech recordings. The aim in this work is to characterize the voice without especially targeting linguistic or prosodic content of the speech. In Study I two groups of methods were tested with vocal tract features outperforming voice source features. The KARMA [46] algorithm was used in Study I and III to extract vocal tract features based on the first three formants. KARMA combines autoregressive (AR) and moving average (MA) modeling of the speech signal by using a Kalman filter. The AR model assumes the lth sample is correlated with the previous p samples and the MA model assumes correlation between the jth sample and noise terms in the previous q samples. Following the windowing, framing and pre-emphasis steps the preprocessed speech signal $\hat{\mathbf{s}}_t[l]$ becomes;

$$s_t[l] = \sum_{i=1}^{p} a_i \hat{\mathbf{s}}_t[l-i] + \sum_{j=1}^{q} b_j u[l-j] + u[l], \tag{3.1}$$

where a_i are the p AR coefficients, b_j are the q MA coefficients and $u[l]$ is the excitation waveform. The z-domain transfer function is then,

$$H(z) = \frac{S(z)}{U(z)} = \frac{1 - \sum_{j=1}^{q} b_j z^{-j}}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{3.2}$$

where $S(z)$ and $U(z)$ are the z-transform of the speech signal $s[l]$ (output) and the excitation signal u[l] (input).

In the KARMA approach the spectral coefficients from Equation (3.2) are transformed to complex cepstrum letting C_n denote the nth cepstral coefficient,

$$C_n = c_n - c_n^{'} \tag{3.3}$$

and the recursive relationship,

$$c_n = \begin{cases} a_n & if\ n = 1 \\ a_n + \sum_{i=1}^{n-1} (\frac{i}{n}) a_{n-i} c_i & if\ 1 < n \leq p \\ \sum_{i=n-p}^{n-1} (\frac{i}{n}) a_{n-i} c_i & if\ n < p \end{cases}$$

$$c_n^{'} = \begin{cases} b_n & if\ n = 1 \\ b_n + \sum_{j=1}^{n-1} (\frac{j}{n}) b_{n-j} c_j^{'} & if\ 1 < n \leq q \\ \sum_{j=n-q}^{n-1} (\frac{j}{n}) b_{n-j} c_j^{'} & if\ n < q \end{cases}$$

ensures the separate contributions to C_n from the denominator and numerator of the transfer function.

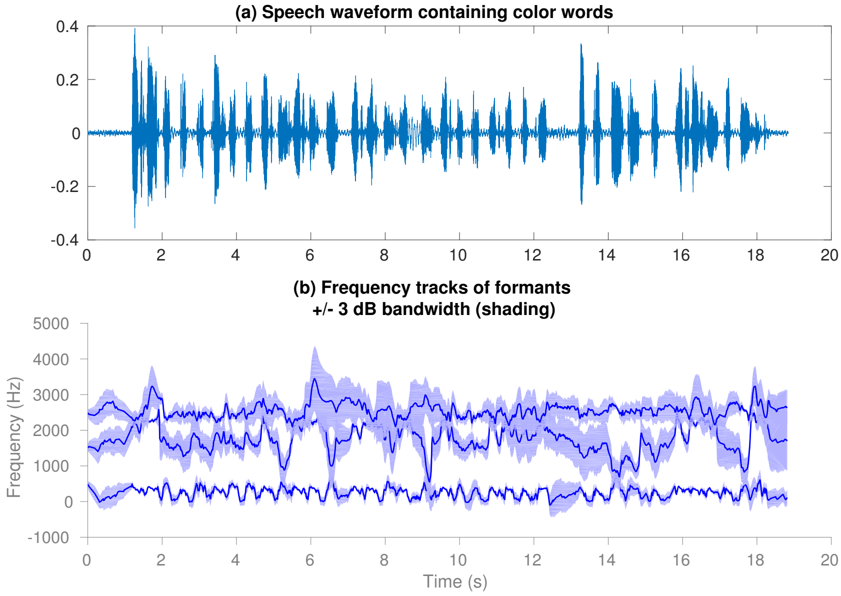


Figure 3.3: An example of one screen from the validation database containing speech waveforms (a) for the utterance of 36 color words. (b) depicts the three frequency track formants for the same screen and their bandwidth using the KARMA algorithm.

Fig. 3.3b illustrates an example of the output of the KARMA algorithm for the speech signal in Fig. 3.3a.

The KARMA algorithm tracks point estimates for formants frame to frame using Kalman inference for cepstrum parameter tracking (for detail see [46]). The main advantage of using KARMA is that the algorithm produces smoother formant tracks than other methods and it provides a sensible interpolation during non-voiced periods.

The formants are directly related to the articulation of the vocal tract and can therefore be interpreted directly. The algorithm was configured to extract three formants, three anti-formants and bandwidths every 10 ms. Three formant frequencies $f_{1,l}, f_{2,l}, f_{3,l}$ were used to produce the feature vector

$$\mathbf{f}_l = [f_{1,l}, f_{2,l}, f_{3,l}]^T \tag{3.4}$$

for the l-th frame in the speech segment (screen) and an $L_j \times 3$ data matrix $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_{L_j}]^T$ where L_j is the total number of frames in the segment.

3.2.2 Cardiovascular measures

The following ten measures for each heartbeat n were obtained from the output of the Finometer Pro system and use in Studies II, III and IV;

- **Heart rate (HR) bpm.,** the most widely used indicator of the status of the cardiovascular system, measured in number of heartbeats pr. minute.

- **Systolic blood pressure (SYS) mmHg.,** the maximum blood pressure during one heartbeat.

- **Diastolic blood pressure (DIA) mmHg.,** the blood pressure during the relaxing period of the cardiac cycle. Both SYS and DIA are measured in millimeters of mercury (mmHg) above the surrounding atmospheric pressure.

- **Mean arterial pressure (MAP) mmHg.,** the weighted sum of SYS and DIA blood pressure calculated as follows:

  $$\mathbf{MAP} = \frac{2}{3}\, DIA + \frac{1}{3}\, SYS \tag{3.5}$$

  MAP represents the passage of fluid through the circulatory system depending on blood pressure and the resistance to flow presented by the blood vessels. The mean blood pressure decreases as the circulating blood moves away from the heart.

- **Stroke volume (SV) mL.,** the volume of blood pumped from the left ventricle per heartbeat. SV is calculated by subtracting end-systolic volume (ESV) from end-diastolic volume (EDV):

  $$\mathbf{SV} = EDV - ESV \tag{3.6}$$

- **Left ventricular ejection time (LVET) ms.,** measures the period of blood flow across the aortic valve. It is influenced by HR, pre-load, after-load, and contractile state [99]. It gives a measurement of the time between upstroke and the dicrotic notch with a normal value of 0.35 +/- 0.08 ms.

- **Pulse interval (PI) ms.,** also known as inter beat interval, is a measurement of the time laps between two consecutive pulsations and its inverse is heart rate [98].

- **Maximum slope (MS) mmHg/s.**, maximum slope of unprocessed pressure rise during the upstroke [98].

- **Cardiac output (CO) l/m.** The volume of blood pumped by the heart per minute.

  $$\mathbf{CO} = SV * HR \tag{3.7}$$

- **Total peripheral resistance (TPR) mmHg.m/l & dyn.s/cm^5**, is a measurement of the resistance that must be overcome to push blood through the circulatory system and create blood flow.

  $$\mathbf{TPR} = \frac{80(MAP - MRAP)}{CO} \tag{3.8}$$

  where MRAP = mean right atrial pressure (in mmHg.), is the average pressure of blood as it returns to the heart.

The cardiovascular measures are designated as $c_{n,i}$ where $i \in \{1, 2, \ldots, 10\}$ is the index for the measure and n is the integer time index of the heartbeat. The ten dimensional feature vector for the n-th heartbeat is therefore,

$$\mathbf{c}_n = [c_{n,1}, c_{n,2}, \ldots, c_{n,10}]^T \tag{3.9}$$

and a $N_j \times 10$ data matrix $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{N_j}]^T$ where N_j is the total number of heartbeats in the segment.

3.2.3 Temporal feature extraction methods

Both psychophysiological feature sets, cardiovascular features $\mathbf{c}_n$ and the formant features $\mathbf{f}_l$, were processed further in order to enrich their representation with a dynamical context. A temporal correlation method was represented in Study I (see Section 4) but later discarded for a method better suited to the type of feature streams used in the following studies. A richer dynamical context was achieved by calculating the so called delta and acceleration features on the static cardiovascular vector $\mathbf{c}_n$ and the formant feature vector $\mathbf{f}_l$ to include information about the rate of change. For an arbitrary feature vector $\mathbf{v}_r$ whose feature index is denoted with i and time index with r the coefficients of the delta vector δ_r are calculated from the measure $v_{r,i}$ along the time index i with

$$\delta_{r,i} = \frac{\sum_{m=1}^{M} m(v_{r+m,i} - v_{r-m,i})}{2\sum_{m=1}^{M} m^2}, \tag{3.10}$$

where M denotes the number of adjacent features before and after r used to derive the delta features. This was set to $M = 2$ for this work (resulting in a window size of 5). The delta vector was therefore obtained by $\delta_r = [\delta_{r,1}, \delta_{r,2}, \ldots]^T$ and the acceleration feature vector $\mathbf{a}_r$ is calculated using the same formula but using $\delta_{r,i}$ instead of $v_{r,i}$. The delta and acceleration vectors were then appended to the feature vector to produce $\mathbf{v}_r^{(\delta,a)} = [\mathbf{v}_r^T, \delta_r^T, \mathbf{a}_r^T]^T$. Hence, a delta and acceleration extended cardiovascular feature vector would be denoted $\mathbf{c}_l^{(\delta,a)} = [\mathbf{c}_n^T, \delta_n^T, \mathbf{a}_n^T]^T$ and its corresponding data matrix $\mathbf{C}^{(\delta,a)}$; for the formant feature vector the extension is $\mathbf{f}_l^{(\delta,a)} = [\mathbf{f}_l^T, \delta_l^T, \mathbf{a}_l^T]^T$ and its corresponding data matrix $\mathbf{F}^{(\delta,a)}$.

3.2.4 Classifier training and evaluation

Two types of supervised learning classifiers were implemented using the statistics toolbox in Matlab: SVM, $\mathbf{y}_{SVM}(\mathbf{x}_n)$ and RF, $\mathbf{y}_{RF}(\mathbf{x}_n)$. The SVM is fundamentally a binary classifier and therefore to solve the trinary classification problem, three two-class one-v.s.-rest binary SVM classifiers were implemented, one for each Stroop level. The soft score output $y_k(\mathbf{x}_n)$ is the signed distance from the decision boundary for each of the three classifiers where $k \in \{1, 2, 3\}$. The class is then determined by the one-v.s.-all classifier which obtains the maximum signed distance from the decision boundary. If all scores are negative then the class closest to the decision boundary with the least negative score is chosen. The training was done using the default linear kernel function and the predictors were standardized before training.

The RF classifier was trained for each heartbeat using one hundred decision trees and the minimum number of observations in a leaf was set to one. The soft score $y_k(\mathbf{x}_n)$ for the random forest classifier is the proportion of trees in the ensemble predicting class k and is interpreted as the probability of this observation $\mathbf{x}_n$ originating from this class.

Throughout this work, participant dependent classifiers were trained where P separate classifiers were trained and tested, one for each participant. A leave-one-out strategy was used where the test sample that is left out corresponds to a single screen j_p. The other twenty screens from that participant were used to train the classifiers. In the case of single heartbeat classifiers the number of test results corresponded to the number of heartbeats contained within the test screen (an average of 37) but in the case of the sequence classification for the entire screen only one result was obtained. The experiment was then repeated with another screen from the set of twenty-one was reserved for testing repeating the procedure 20 times. This procedure was then repeated for each participant in the study.

3.2.5 Classifier design

All of the classifiers were evaluated with two types of CE, either (1) from a single heartbeat feature vector $\mathbf{x}_n$ or (2) a sequence of heartbeats from a single screen represented with the data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]^T$. The classifiers soft (likelihood) scores were retrieved and are denoted either as $y_k(\mathbf{x}_n)$ for the single heartbeat classifier or $y_k(\mathbf{X})$ for the sequence classifier. The index $k \in \{1, 2, 3\}$ denotes Stroop level one, two or three respectively. The vector,

$$\mathbf{y}(\mathbf{x}_n) = [y_1(\mathbf{x}_n), y_2(\mathbf{x}_n), y_3(\mathbf{x}_n)]^T \tag{3.11}$$

contains the soft scores for the heartbeat at time index n and the heartbeat classification simply chooses the class with the maximum value in that vector for each n. For the sequence classification, the soft scores are collected in an output matrix

$$\mathbf{Y_k} = [\mathbf{y_k}(\mathbf{x}_1), \mathbf{y_k}(\mathbf{x}_2), \dots, \mathbf{y_k}(\mathbf{x}_N)]^T \tag{3.12}$$

for each $k \in \{1, 2, 3\}$. The classification result is then obtained by summing the soft scores together for each k over the screen and concatenating them to obtain a single score for each class

$$\mathbf{y}(\mathbf{X}) = [y_1(\mathbf{X}), y_2(\mathbf{X}), y_3(\mathbf{X})]^T \tag{3.13}$$

and then the class according to the maximum value of that vector is chosen.

Part II

Studies

Chapter 4

Study I Vocal tract and voice source features

The first study focuses on analyzing the speech data with the objective of *providing an independent verification of whether there is a link between increased cognitive workload and changes in the speech signal and, if this link exists, to characterize what part of voice is mostly affected.* From the outset the objective was to investigate the feasibility of using features extracted from the voice to indicate cognitive workload levels. This objective was also supported by research indicating a strong relationship between cognitive workload and features extracted from speech signals [6], [66].

The methods introduced for speech processing in Study I were based on two groups of feature extraction methods; vocal tract and voice source feature streams. These feature streams were fused at different stages of the process; at the feature level, the utterance level and the output level. A temporal correlation structure was calculated for the two feature streams and their concatenated form (feature level fusion). These feature streams fused at different levels were finally tested on a participant dependent basis, with three sets of classifiers; minimum distance, SVM and RF. Fig. 4.1 depicts the steps included in the processing flow and at which level the different fusion schemes took place. By applying weights to the two feature streams, vocal tract and voice source features at the output level the optimum MCR of 32.5% was reached with the vocal tract features (weight= 0.76) dominating the voice source features (weight= 0.24).

4.1 Methods

The approach adopted in this work was based on the idea that decomposing the speech signal into the vocal tract characteristics and the voice source signal should give insight into the voice changes brought by increased cognitive workload. Two sets of distinct features were extracted from the speech signal, the first set consisted of features describing the vocal tract and the second set consisted of features describing the voice source signal. These groups of features were chosen following the results published by the NICTA group [6]. Here the focus is on researching their compatibility with a temporal correlation structure and different classification methods for cognitive workload detection.

4.1.1 Vocal tract feature extraction

The set of vocal tract features used in this work consisted of ordinary formants f_1, f_2 and f_3. These were extracted from the speech signal using the KARMA [46] algorithm (see Section 3.2.1). The three formants were obtained for each 20 ms of speech- and concatenated in a vocal tract feature vector $\mathbf{x}_{vt}(j)$, where j denotes the frame index.

4.1.2 Voice source features

The voice source features consisted of 10 different parameters that were extracted either directly from the speech signal, from an estimate of the glottal flow or its derivative (voice source signal). The following is a list of the parameters:

- **Fundamental frequency ($\mathbf{F_0}$),** the lowest frequency of a harmonic series, is a measure of the vibration of the vocal cords [43].
- **Harmonic richness factor (HRF) and $\mathbf{1^{st}}$ to $\mathbf{2^{nd}}$ harmonics ratio (H1-H2),** are both measures of spectral descriptors of the single pulse shape extracted from the true glottal flow [100], [101].
- **Harmonics to noise ratio (HNR),** represents the degree of periodicity [102].
- **Cepstral peak prominence (CPP),** has been applied as an acoustic measure of voice quality that measures the degree of harmony within a voice sample [103].
- **Normalized amplitude quotient (NAQ),** the peak-to-peak glottal flow amplitude is divided by the product of fundamental period and maximum flow declination rate for each cycle of the true signal [104].
- **Pulse amplitude (PA),** is together with NAQ an example of pulse shape methods used in glottal source analysis.
- **Jitter,** represents the small fluctuations in glottal cycle lengths [105].
- **Closed quotient (CQ),** is defined as the percentage of each cycle for which the vocal folds are in contact [106].
- **Maximum flow declination rate (MFDR),** measures how rapidly the vocal folds are closing and correlates with vocal intensity [107].

Cepstral peak prominence (CPP) was the only measure extracted directly from the speech signal and has been widely used to classify and rate levels of dysphonia [103]. While there is no clear understanding of what the parameter measures, the general findings among the speech pathologists show that the parameter is tied to vocal attributes of breathiness, roughness and hoarseness [108]. A theoretical study on the parameter's nature was done in [108] where the authors concluded that CPP integrates measure of several features describing the aperiodicity and waveform of the acoustic voice signal.

The iterative adaptive inverse filtering algorithm [104] was employed in order to obtain the voice source signal from which the rest of the voice source measures were extracted. Authors in [109] studied the relation of these features to five emotions (joy, anger, fear, sadness and relief) and found a statistical significance for all of them except NAQ. However, other works [110] have reported on increase of this parameter for hypo-functioning (e.g. relived) voice and a decrease for hyper-functioning voice (e.g. angered) [111]. The voice source features were obtained for each 20 ms of speech- and concatenated in a voice source feature vector $\mathbf{x}_{vs}(j)$, where j denotes the frame index.

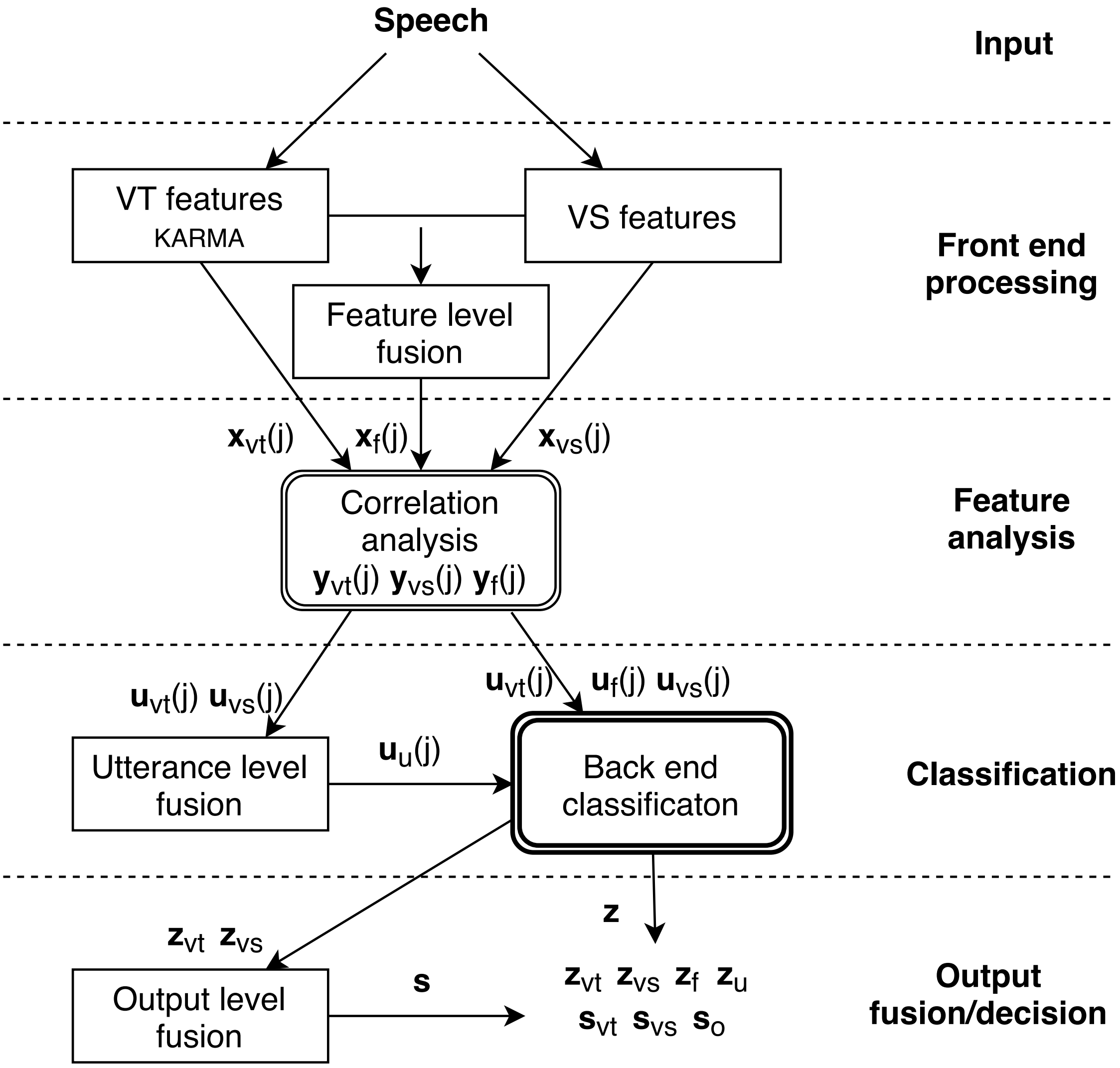

Speech
Input
VT features
KARMA
VS features
Feature level fusion
Front end processing
$\mathbf{x}_{\mathrm{vt}}(j)$
$\mathbf{x}_{\mathrm{f}}(j)$
$\mathbf{x}_{\mathrm{vs}}(j)$
Correlation analysis
$\mathbf{y}_{\mathrm{vt}}(j)$ $\mathbf{y}_{\mathrm{vs}}(j)$ $\mathbf{y}_{\mathrm{f}}(j)$
Feature analysis
$\mathbf{u}_{\mathrm{vt}}(j)$ $\mathbf{u}_{\mathrm{vs}}(j)$
$\mathbf{u}_{\mathrm{vt}}(j)$ $\mathbf{u}_{\mathrm{f}}(j)$ $\mathbf{u}_{\mathrm{vs}}(j)$
Utterance level fusion
$\mathbf{u}_{\mathrm{u}}(j)$
Back end classificaton
Classification
$\mathbf{z}$
$\mathbf{z}_{\mathrm{vt}}$ $\mathbf{z}_{\mathrm{vs}}$
Output level fusion
$\mathbf{s}$
$\mathbf{z}_{\mathrm{vt}}$ $\mathbf{z}_{\mathrm{vs}}$ $\mathbf{z}_{\mathrm{f}}$ $\mathbf{z}_{\mathrm{u}}$
$\mathbf{s}_{\mathrm{vt}}$ $\mathbf{s}_{\mathrm{vs}}$ $\mathbf{s}_{\mathrm{o}}$
Output fusion/decision

Figure 4.1: Flow chart describing the steps of the whole process from front end processing to output decision. The fusion levels are depicted at the levels they are performed with their relevant symbols.

4.1.3 Feature level fusion

Feature level fusion was achieved by concatenating the frame-level feature vectors of the vocal tract features ($n_{c-vt} = 3$ formant values) and voice source features ($n_{c-vs} = 10$ parameters) for each frame, so that, instead of having only $\mathbf{x}_{vt}(j)$ or $\mathbf{x}_{vs}(j)$, we now have $\mathbf{x}_f(j) = [\mathbf{x}_{vt}(j)^T, \mathbf{x}_{vs}(j)^T]^T$ for each frame j. From this a correlation matrix (with $n_c = 13$ parameters) and an utterance-level feature vector $\mathbf{u}$ was made.

4.1.4 Temporal correlation structure

The three feature streams $\mathbf{x}_{vt}(j)$, $\mathbf{x}_{vs}(j)$ and $\mathbf{x}_f(j)$ were processed further and summarized to produce the feature vectors $\mathbf{u}_{vt}$, $\mathbf{u}_{vs}$ or $\mathbf{u}_f$ for the utterance level classification. This was done by calculating the correlation structure of the feature stream as introduced in [112]. A fixed time scale of 2 frames was used to create a concatenated feature vector of the current one at time j with 13 successive time delays. For the joint vocal tract and voice source vector $\mathbf{x}_f(j)$ the new data vector of $N = n_c n_d = 13 \times 14 = 182$ dimensions was created with,

$$\mathbf{y}_f(j) = [\mathbf{x}_f^T(j), \mathbf{x}_f^T(j-2), \mathbf{x}_f^T(j-4), \ldots \mathbf{x}_f^T(j-26)]^T. \tag{4.1}$$

By applying the same concatenation method, the vocal tract $\mathbf{y}_{vt}(j)$ and voice source feature vectors $\mathbf{y}_{vs}(j)$ became $N_{vt} = n_{c-vt} n_d = 3 \times 14 = 42$ and $N_{vs} = n_{c-vs} n_d = 10 \times 14 = 140$ dimensions respectively.

The cross-correlation matrix for the utterance was then formed using,

$$\mathbf{R} = \frac{1}{N_s} \sum_j \tilde{\mathbf{y}}(j) \tilde{\mathbf{y}}^T(j) \tag{4.2}$$

where N_s is the number of frames in the utterance and $\mathbf{y}(j)$ is normalized to have a zero mean and unit variance, denoted as $\tilde{\mathbf{y}}(j)$. The eigenvalues λ_i were obtained from $\mathbf{R}$ in descending order and each component in the utterance level feature vector $\mathbf{u} = [u_1, u_2, \ldots, u_N]^T$ was then computed as the normalized cumulative sum

$$u_n = \frac{\sum_{i=1}^{n} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \tag{4.3}$$

where $n = 1, 2, \ldots, N$. The component u_n represents the proportion of energy contained in the first n eigenvectors.

Fig. 4.2 illustrates the extracted vocal tract $\mathbf{u}_{vt}$ and voice source $\mathbf{u}_{vs}$ feature vectors for three cognitive workload levels. The figure demonstrates that for these utterances with the vocal tract feature stream, more than 90% of the energy is captured by the first 8 out of $N_{vt} = 42$ components. Whereas for the voice source features more than 43 out of $N_{vs} = 140$ components are needed to capture 90% of the energy. The number of eigenvalue components in the figure are truncated for illustrative purposes after the cumulated eigenvalues reached 99.9%.

4.1.5 Utterance level fusion

Utterance level fusion was achieved by concatenating the high-level feature vectors $\mathbf{u}_{vt}$ and $\mathbf{u}_{vs}$ into a single vector $\mathbf{u}_u$. This approach is distinctly different from the frame-level fusion. Feature-level fusion assumes that modeling cross-correlations between vocal tract and voice parameters might provide new information about the cognitive workload. Utterance-level fusion, on the other hand, assumes that these two streams are independent of each other and thus can be used in tandem.

4.1.6 Output level fusion

The output level fusion was performed by accumulating the soft scores outputs from the classifiers for both vocal tract $\mathbf{z}_{vt}$ and voice source $\mathbf{z}_{vs}$ feature streams into a

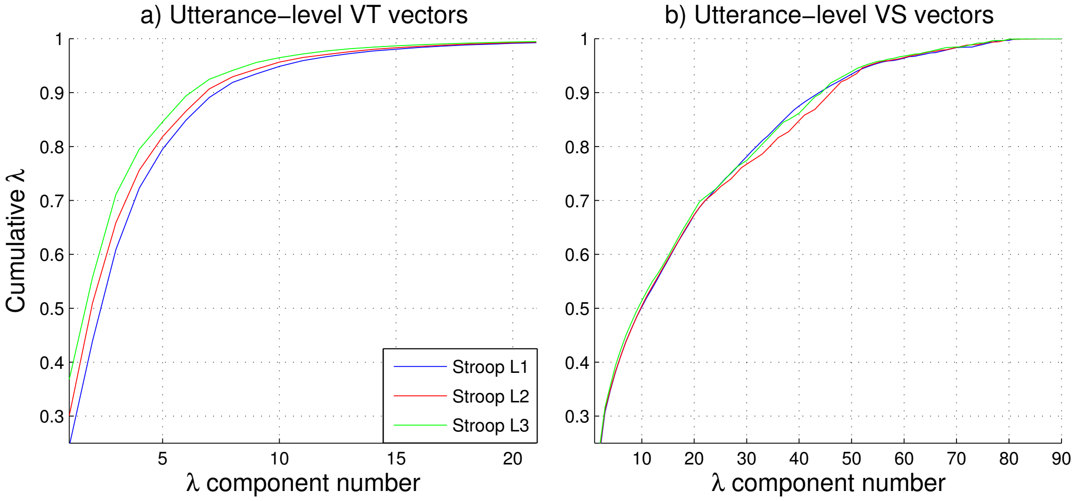


Figure 4.2: Example of the energy contained in the first 99.9% cumulated eigenvalues. a) depicts the energy for vocal tract features and b) for voice source features for all three Stroop tasks.

single score. The weights for particular streams were analyzed with sensitivity analysis in order to determine their optimal values subject to the criteria $\sum_i w_i = 1$. The accumulated scores were then classified using that maximum a posteriori criteria.

Sensitivity analysis was performed to gain insight into whether a weighted combination of the feature streams could outperform the vocal tract feature stream with the SVM classifier fused at the output level. The optimum combination of weights w between $\mathbf{s}_{vs}$ and $\mathbf{s}_{vt}$ feature streams where calculated,

$$\mathbf{s}_o = w\mathbf{s}_{vs} + (1 - w)\mathbf{s}_{vt} \tag{4.4}$$

4.2 Results

The results obtained from the classification tests are shown in Table 4.1. There are several interesting things that can be highlighted. The first thing to look at is the performance of the proposed speech features in the task of classifying the Stroop level and its correlation to the cognitive workload. A trinary classification task with a set of completely random features would theoretically achieve the MCR of 66%, but the results for our best features were below 33%. These results indicate that the cognitive workload causes changes to the voice that can be observed and objectively measured.

Before the output level fusion was applied the vocal tract features achieved better MCR than any other set of features. The overall best results of 33.92% were obtained with the SVM classifier, while the RF and MD classifiers followed with the MCR of 43.15% and 43.93% respectively. Extended further with the optimum combination of weights between $\mathbf{s}_{vt}$ and $\mathbf{s}_{vs}$ occurring in the weights $w = 0.24$ applied to the voice source feature stream, reducing the MCR to 32.5%. The second best results were achieved with combined features (both utterance and feature-level fused) and the voice source features scored as the last. Another interesting thing to note is the fact that utterance-level fusion outperformed the feature-level fusion by 5.49% with the SVM classifier. In fact, in all studied setups the SVM proved to consistently outperform the other two classifiers.

Table 4.1: Average MCR±SE [%] over all participants with different sets of feature streams and classifiers.

Features	**MD**	**SVM**	**RF**
VTfeat $\mathbf{z}_{vt}$	43.93±1.09	**33.92±1.05**	43.15±1.09
VSfeat $\mathbf{z}_{vs}$	65.65±1.05	47.47±1.1	55.59 ± 1.1
Utt. Fused $\mathbf{z}_{u}$	53.43±1.08	35.86±1.06	41.77±1.09
Feat. Fused $\mathbf{z}_{f}$	42.91±1.09	41.35± 1.09	49.08±1.1

A comparison of the output level fusion results is presented in Table 4.2. In the table the results of the SVM classification scheme and the soft score classification are represented. The results clearly show that by applying the output level fusion method the combined feature streams, vocal tract and vocal source, outperform the result of vocal tract feature stream and SVM classification combination already presented.

Table 4.2: Comparison of soft score classification and SVM classification. MCR [%] between vocal tract features and voice source features.

	Soft score	**SVM**
	s	**z**
VT feat	34.49	33.92
VS feat	46.38	47.47
Utt.fusion	-	35.86
Output fusion	**32.50**	-

An insight into the separability of the distinct Stroop levels can be achieved by taking a closer look at a confusion table between the three levels. Table 4.3 presents the results for SVM classifiers with the vocal tract features averaged over all participants. The actual Stroop L1 screens were more often classified as Stroop L2 then they were with the Stroop L3 tasks. The same trend, although naturally reversed, can be observed for the actual Stroop L3 screens. This trend seems much more pronounced for actual Stroop L3 tasks leading to the conclusion that an increase in cognitive workload produces changes in the voice that can be extracted using the methods presented in this work. Another interesting fact is that a classifier trained for Stroop L2 tasks misclassified the Stroop L3 features more often than Stroop L1. This observation leads to the conclusion that the increased cognitive difficulty of Stroop L2 and Stroop L3 tasks introduces changes to the voice, which can be accurately detected using the vocal tract features.

A histogram of MCR for all 98 participants is presented in Fig. 4.3. The figure highlights the extreme difference between individuals classification results. A single participant scored MCR of 60%, which is very close to random classification, while a single participant scored MCR of 0%, which represents a perfect classification score. Even though most of the participants scored around the average MCR of 33% this describes in essence the impact of individuality on generalized classification.

Table 4.3: Confusion table for the Stroop levels showing the average MCR [%] for all participants with the vocal tract features with the SVM classifier.

	Stroop L1	**Stroop L2**	**Stroop L3**
Stroop L1	**72.5**	15.7	11.8
Stroop L2	17.5	**59.7**	22.8
Stroop L3	10.1	24.6	**65.3**

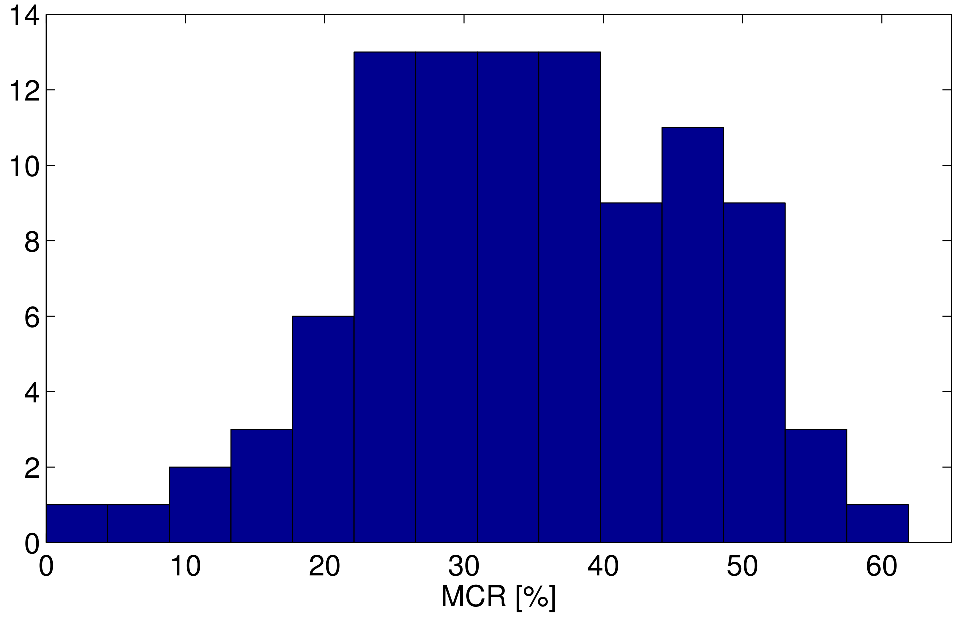


Figure 4.3: A histogram of the MCR [%] for all participants with the results from vocal tract features with the SVM classifier.

4.3 Discussion and Conclusions

This study gives an independent verification of the relationship between cognitive workload and speech. Participant independent classifiers trained with 98 participants were able to distinguish between utterances of low, medium and high cognitive workload with 32.5% MCR with the weighted combination of vocal tract and vocal source features fused at the output level. The vocal tract features achieved 33.92% MCR and the voice source features only achieved 46.38% MCR on their own.

Specific conclusions of this work are that the particular vocal tract parameters used in this work [46] outperform the voice source parameters that were studied in [109]. This is in concurrence with studies that compare vocal tract and voice source features both for other tasks [48], [113] and cognitive workload [41], [63]. More generally, the study reinforces previous findings that show that there is a link between cognitive workload and speech [41], [63], [67], [114]–[116].

An average MCR of over 30% might seem high, given that each test utterance contains information of over 20 s. However, the task of trinary classification is not commonly attempted no matter the measure and the challenge of individuality plays a major role in all generic attempts of classifications.

The fact that cardiovascular signals were recorded simultaneously with the speech signal provided us with the opportunity of investigating methods to approach concerns regarding the time-varying nature of the body's response to cognitive workload. This is indeed the focus of the next research, by investigating the information embedded in the cardiovascular signals and the detectable reactions to cognitive workload.

Chapter 5

Study II Cardiovascular signals with short term dynamic features

In the second study the objective was to *establish whether cardiovascular signals could be used to perform trinary classification of cognitive workload.* Cardiovascular signals have been proven to be a good indication of reactions to cognitive stimuli but researchers have not been able to move beyond binary classification of high/low or normal/high load. These methods have failed to capture the complicated dynamic relationship of the sympathetic and the parasympathetic systems and their role in regulating reactions of the cardiac system (see: 1.4.1 for detail on neural regulation of the cardiac cycle). Here, ten cardiovascular signals representing a hemodynamic profile of the participants are analyzed capturing their short-term dynamic variability with a method known from speech processing. By reducing the segments to two adjacent heartbeats a trinary classification on individual basis is achieved on both screen and heartbeat basis.

5.1 Methods

The feature sets containing cardiovascular signals were constructed with parameter values describing the hemodynamic profile during each heartbeat. Temporal information on the changes occurring in adjoining heartbeats were included for each of these parameter values with delta and acceleration calculations (see: Section 3.2.3). Two distinct supervised learning classifying methods were used and their likelihood score used for the classification schemes of (1) each heartbeat and (2) each task screen. The feature set combinations in this particular work are based on the cardiovascular measures either from a single heartbeat feature vector $\mathbf{x}_t$ (without delta features, $\mathbf{x}_t = \mathbf{c}_t$ or with delta features, $\mathbf{x}_t = \mathbf{d}_t$ or $\mathbf{x}_t = \mathbf{e}_t$) or a sequence of heartbeats from a single screen represented with the data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]^T$. The results show that SVM outperforms RF with the average MCR of $\mathbf{20.44\%}$ using the whole screen classification scheme instead of individual heartbeat classification.

A more detailed description of the methods used in Study II are presented in Section 3.2.2. Some changes were made to the procedures based on the results and experience from Study I such as excluding the MD classifier. The nature of the feature set made the temporal correlation structure unfeasible therefore the dynamic information was captured using a method known in the speech processing field called delta and acceleration(see Section 3.2.3 for detail). The conclusion of this study was that a finer cognitive workload distinction can be reached with the combined feature set of cardiovascular signals and delta coefficients, classifying for each screen with the SVM classifier.

5.2 Results

The results of the study are first presented for the SVM classifier with the temporal delta feature sets. Then a comparison of the results from single heartbeat classifier and sequence screen classifier is reported. The results are then summarized as single average MCR and compared with the other feature set combinations as well as the results of the RF classifier. Finally, a further analysis of the performance of individual participants is presented.

5.2.1 Support vector machines using temporal features

Table 5.1 lists a confusion table of how all heartbeats in the data set were classified using the SVM classifier $\mathbf{y}_{SVM}(\mathbf{x}_t)$ and the delta-delta feature set $\mathbf{e}_t$. The numbers are obtained by summing individual confusion tables of all the participants.

There misclassification, seems to be evenly distributed over the remaining classes which indicates that no level is dominating the classification. However, Stroop level 3 seem to be performing slightly better than the other levels with a 23.13% MCR and 25.69% MTR. The participant-average test set MCR is given as 29.26%. Table 5.2 shows the confusion table for the classification for screens also using the SVM $\mathbf{y}_{SVM}(\mathbf{X})$ and the temporal delta-delta feature set $\mathbf{e}_t$. As with the heartbeat confusion table, the numbers are obtained by summing individual confusion tables of participants. In both tables the MCR and MTR are also listed for the Stroop levels as well as the participant-average test set MCR. The table shows that the screen based classifier is also balanced with respect to the Stroop levels.

Table 5.1: Confusion table for the number of heartbeats classified with the $\mathbf{e}_t$ feature set classifying with the SVM $\mathbf{y}_{SVM}(\mathbf{x}_t)$.

	Stroop L1	**Stroop L2**	**Stroop L3**	**MCR [%]**
Stroop L1	**13790**	3749	3554	34.62
Stroop L2	3745	**16536**	4142	32.29
Stroop L3	3022	3676	**22254**	23.13
MTR [%]	32.92	30.99	25.69	**29.26**

The advantage of using a sequence of heartbeats is clearly seen by comparing Table 5.1 and Table 5.2 as participant-average test set MCR improves from 29.26% to 20.44%.

5.2.2 Overall performance results

Table 5.3 compares the feature sets by listing the participant-average test set MCR for the SVM classifier. The rows list the results for the three feature sets: without delta

Table 5.2: Confusion table for the number of screens classified with the $\mathbf{e}_t$ feature set classifying with the SVM $\mathbf{y}_{SVM}(\mathbf{X})$.

	Stroop	**Stroop**	**Stroop**	**MCR**
	L1	**L2**	**L3**	**[%]**
Stroop L1	**499**	95	78	25.74
Stroop L2	63	**440**	73	23.61
Stroop L3	38	65	**665**	13.41
MTR [%]	16.83	26.66	18.50	**20.44**

features, $\mathbf{c}_t$, with delta features, $\mathbf{d}_t$ and with delta-delta features $\mathbf{e}_t$. The two columns list the results for the screen sequence classifier $\mathbf{y}_{SVM}(\mathbf{X})$ and the heartbeat classifier $\mathbf{y}_{SVM}(\mathbf{x}_t)$. Each result quantifies the average MCR along with the SE. For example the SVM screen classifier using the cardiovascular feature set without delta features, $\mathbf{c}_t$, obtains average MCR of 26.34% ± 16.95%.

Table 5.3: Average MCR±SE [%] over all participants comparing different combinations of classification schemes and feature sets for the SVM classifier.

Feature set	**Screen**	**Heartbeat**
$\mathbf{x}_t$	$\mathbf{y}_{SVM}(\mathbf{x}_t)$	$\mathbf{y}_{SVM}(\mathbf{X})$
$\mathbf{c}_t$	26.34±16.95	36.11±13.88
$\mathbf{d}_t$	23.42±17.08	33.73±14.11
$\mathbf{e}_t$	**20.44±15.48**	29.26±13.27

The table shows that adding delta-delta features improves performance considerably for both the screen based and the heartbeat based SVM classifier. The best results of 20.44 ± 15.48% were obtained using delta and delta-delta features, $\mathbf{e}_t$ and the screen based classifier $\mathbf{y}_{SVM}(\mathbf{X})$. These results are also the best results obtained in this study and correspond to the confusion table presented in Table 5.2. Further analysis of these results are presented in Section 5.2.3.

Table 5.4 depicts the set of results for the RF classifier, were the overall results are inferior to the SVM results. For the heartbeat based RF classifier the performance improves by adding delta-delta features which is consistent with the SVM results. However the screen based classifier is not improved by adding the delta-delta features.

Table 5.4: Average MCR±SE [%] over all participants comparing different combinations of classification schemes and feature sets for the RF classifier.

Feature set	**Screen**	**Heartbeat**
$\mathbf{x}_t$	$\mathbf{y}_{RF}(\mathbf{X})$	$\mathbf{y}_{RF}(\mathbf{x}_t)$
$\mathbf{c}_t$	22.72±15.99	37.21±12.79
$\mathbf{d}_t$	23.91±15.69	35.75±12.73
$\mathbf{e}_t$	24.21±15.89	34.50±12.76

5.2.3 Participant distributions

Fig. 5.1 shows a histogram of the test set MCR for all participants for the SVM classifier with the delta-delta feature set. The first panel contains the results for the single heartbeat classifier $\mathbf{y}_{SVM}(\mathbf{x}_t)$ and the second panel the screen based classifier $\mathbf{y}_{SVM}(\mathbf{X})$. The figure corresponds to the last line in Table 5.3 and demonstrates how the results improve when the classifier can accumulate the results over an entire screen before making a decision. The figure also illustrates how widely distributed the results are depending on the participants, as is evident in the standard deviation. The range for the screen based classifier is from zero MCR to 76.19%. More interesting results evident from the histogram is that 30 participants (out of 96) achieve MCR under 10%.

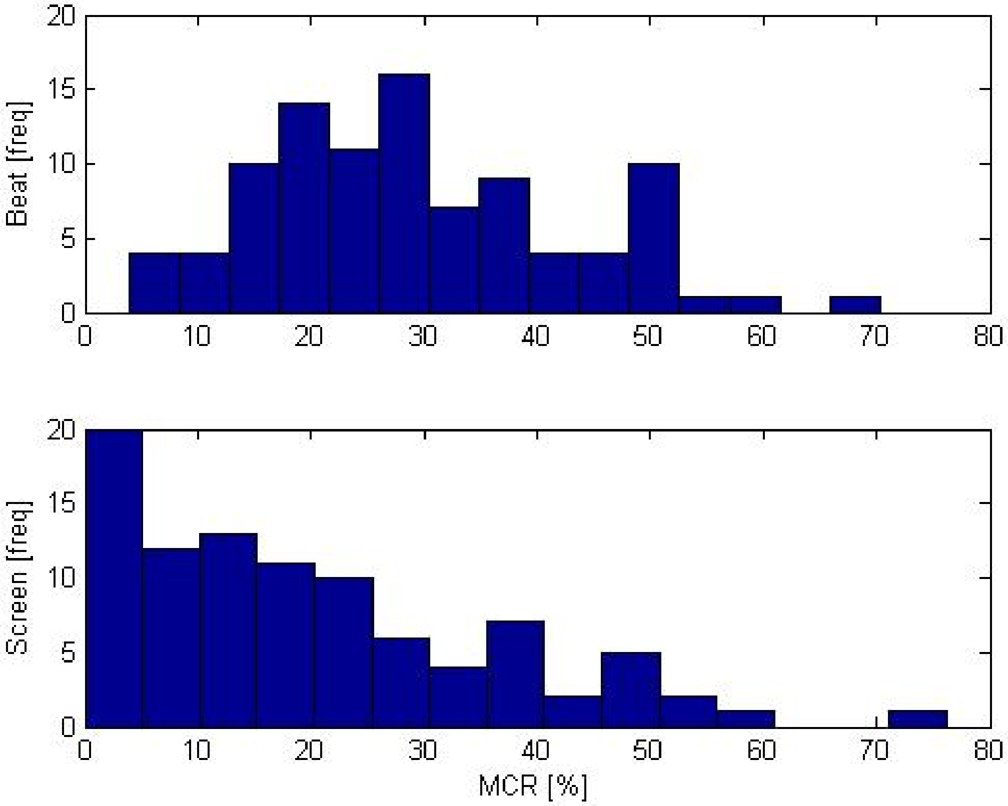


Figure 5.1: Histogram of test set MCR [%] for all participants in the set using the SVM classifier with the delta and delta-delta features. The upper panel shows the individual heartbeat classifier $\mathbf{y}_{SVM}(\mathbf{x}_t)$ results and the lower panel shows the screen-based classifier $\mathbf{y}_{SVM}(\mathbf{X})$ results.

Fig. 5.2 shows a histogram of the MCR with the RF classifier without dynamic features. The figure illustrates how well the performance improves when using a sequence of heartbeats to do the classification. The average MCR for the screen-based classification is 22.72% and the range if from zero to 85.71% (not shown in the figure). The number of participants achieving MCR under 10% falls to 26 participants (out of 96).

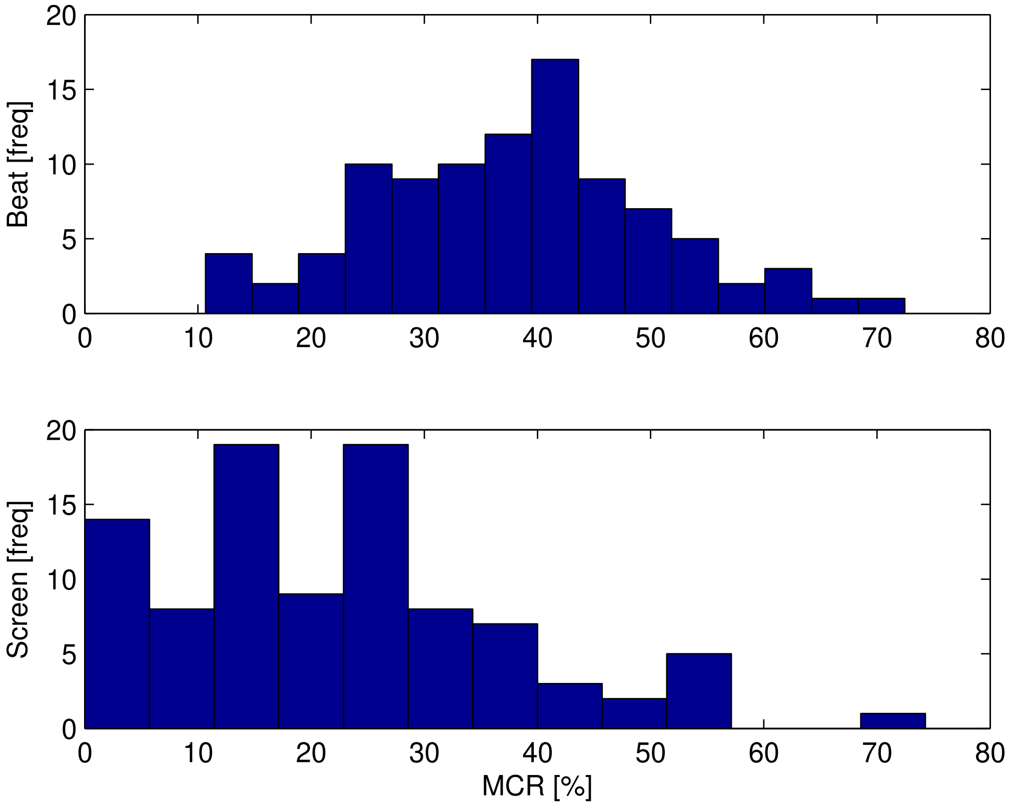


Figure 5.2: Histogram of test set MCR [%] for all participants in the set using the RF classifier without the dynamic features. The upper panel shows the individual heartbeat classifier $\mathbf{y}_{RF}(\mathbf{X})$ results and the lower panel shows the screen-based classifier $\mathbf{y}_{RF}(\mathbf{x}_t)$ results.

5.3 Discussion and Conclusions

The experiments presented in this work depict a method for detecting the cardiovascular systems responds to cognitive stimuli during Stroop tests. The results show obvious cardiovascular reactivity to the different cognitive stimuli. The best classifier was able to distinguish between low, medium and high Stroop levels with an average of 20.44% test set MCR. Further analysis shows that more than 34 participants out of 96 achieve 10% MCR or less. This indicates that cognitive workload strongly affects the cardiovascular system but also that these affects are highly various between individuals. The methodology does not however give an insight into what aspect of the cardiovascular measures are affected or the potential cognitive workload profiles of the individuals according to their susceptibility to the process.

Comparing the performance of the two classifiers the SVM outperformed RF in all instances. The most prominent SVM results achieved 20.44% MCR while the best RF result is 22.72% MCR. Both methods benefit greatly from classifying whole sequence of heartbeats (screens) instead of individual heartbeats. The work also introduced temporal feature extraction methods suited for the cardiovascular signal. Classification with bot classifiers (SVM and RF) for CE of single heartbeats benefited from the addition of these temporal features, however the screen-based RF classifier did not benefit from the addition.

Signal classification offers a new approach to cognitive workload monitoring. The results presented in this study compare well with other classification work in this field. For most parts, reported results are based on a binary classification, sometimes using no vs. some or low vs. high cognitive workload [5], [27], [57], [80], [117]. It is clear from prior work, that high accuracy in binary classification is possible, in particular, if combining multiple physiological signals. The obstacle however, has been to move beyond the binary classification.

The present study goes beyond the state-of-the-art by successfully demonstrating a trinary classification of low, medium and high cognitive workload states. In Study I, three workload states were classified using vocal tract features and SVM with the MCR of 33.5%. The present results achieved a better classification accuracy using cardiovascular signals. As a consequence of these findings, continuing research should aim to explore the synergistic affects of multi-modal measurement techniques combining cardiovascular and audio signals performing the same classification tasks.

In conclusion, cognitive workload classification based on the cardiovascular signal is possible and might provide a reliable, non-intrusive monitoring tool.

Chapter 6

Study III Heartbeat synchronized psychophysiological features

The objective of the third study was to *combine the two psychophysiolgical signals in an attempt to see if they can compliment each other to a better trinary classification results.* By achieving this, a step might be taken towards greater granularity in the task of classifying cognitive workload. To our knowledge, no research has focused on combining speech- and cardiovascular signals for cognitive workload classification. However, much effort has been put into researching different combinations of psychophysiological measures such as EEG and cardiovascular signals in conjunction with other signals or measures. For the trinary classification scheme (low, medium, high cognitive workload) the prominent result achieved 15.17±0.79% average MCR±SE indicating good discrimination at three levels of cognitive workload.

6.1 Methods

The feature extraction methods for the cardiovascular and speech signals applied in this study are described in Chapter 3.

The cardiovascular features $\mathbf{c}_n$ and the formant features $\mathbf{f}_l$ were processed to enrich the representation with a dynamical context and to fuse these two information sources together for a joint classification. Fig. 6.2 depicts the feature extraction, synchronization process, feature set combinations and classification schemes introduced. A richer dynamical context was achieved by calculating the delta-delta features (Section 3.2.3) and a feature level fusion is attained by calculating formant statistics around each heartbeat. This is outlined below.

6.1.1 Synchronizing formant features and heartbeats

Each audio recording of a single screen j containing N_j heartbeats and L_j formant frames. The heartbeats occur at τ_n seconds into the recordings, and each pulse interval is therefore $\Delta_n = \tau_n - \tau_{n-1}$. The start of each formant frame occurs at t_l and the fixed formant frame increment is $t_l - t_{l-1} = 10$ ms. A typical pulse interval is $\Delta_n = 700$ ms (c.a. 86 beats-per-minute) which gives a ratio of 70:1 formant frames to heartbeat. Fig. 6.1 illustrates the synchronization process graphically with heartbeats (stem) and the vertical dotted lines mark the number of formant frames attached to it.

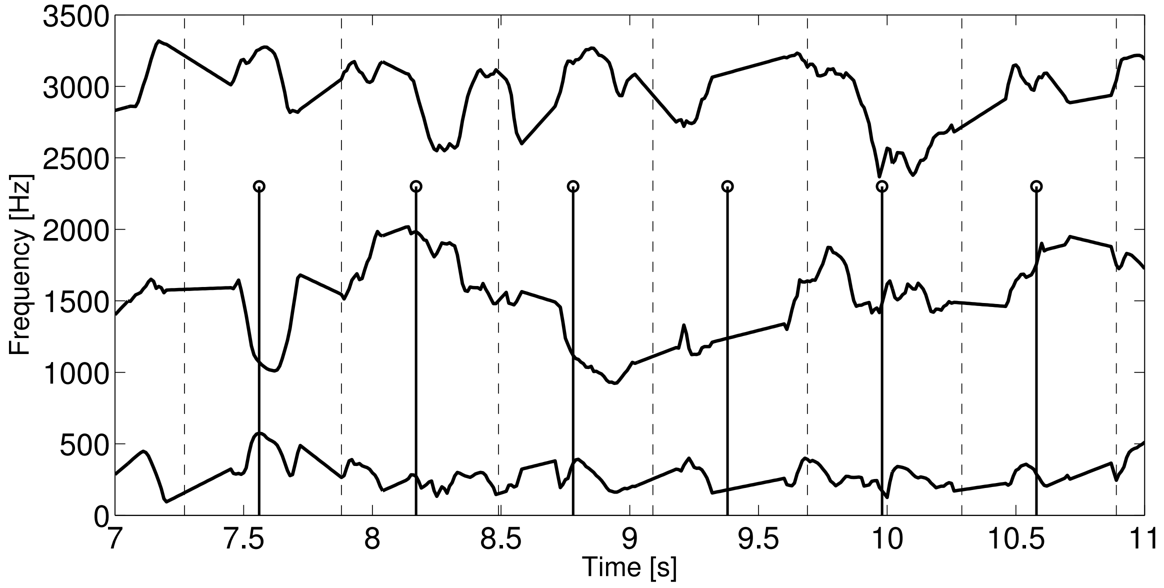


Figure 6.1: An example of the synchronization process with an approximately 10 heartbeat portion of a screen. The plot shows on a disproportional y-axis the three formant tracks. The heartbeats are indicated by the stem lines and the vertical dotted lines represent the interval on either side of the heartbeat marking the formant frames assigned to each heartbeat.

The formant feature data matrix $\mathbf{F}$ was divided into N_j sub-matrices $\mathbf{F}_n$ which correspond to each heartbeat. The matrix $\mathbf{F}_n = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{M_n}]^T$ has the row vector $\mathbf{f}_m$ where $m \in \{l | t_l > \tau_n - \frac{1}{2}\Delta_n, t_l < \tau_n + \frac{1}{2}\Delta_{n+1}\}$. The three columns in $\mathbf{F}_n$ correspond to three formant tracks in the neighborhood of the n-th heartbeat. Ten features were calculated, e.g. for the first formant track, the first three parameters are the coefficients of a fitted second order polynomial, denoted as $\phi_{1,n}$, $\phi_{2,n}$, and $\phi_{3,n}$ and then the minimum $\phi_{4,n}$ and maximum $\phi_{5,n}$ values, average $\phi_{6,n}$, median $\phi_{7,n}$, standard deviation $\phi_{8,n}$, skewness $\phi_{9,n}$ and kurtosis $\phi_{10,n}$. The same features were calculated for the second $\phi_{11-20,n}$ and third formant tracks $\phi_{21-30,n}$. The corresponding feature vector then becomes,

$$\boldsymbol{\phi}_n = [\phi_{1,n}, \phi_{2,n}, \ldots, \phi_{30,n}]^T \tag{6.1}$$

and the $N_j \times 30$ dimensional data matrix for the segment is $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_{N_j}]^T$. Similarly the delta-delta expanded formant data matrix $\mathbf{F}^{(\delta,a)}$ is divided up into sub-matrices $\mathbf{F}_n^{(\delta,a)}$ in the same way to produce three formant tracks, three delta formant tracks and three acceleration formant tracks in the neighborhood of the n-th heartbeat. The same features were calculated for each of the nine tracks to produce a ninety-dimensional feature vector

$$\gamma_n = [\gamma_{1,n}, \gamma_{2,n}, \ldots, \gamma_{90,n}]^T \tag{6.2}$$

and an $N_j \times 90$ dimensional data matrix for the segment $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, ..., \boldsymbol{\gamma}_{N_j}]^T$. Notice that the delta-delta operators are applied before the heartbeat synchronized features are calculated and could be reapplied again to produce the data matrices $\boldsymbol{\Phi}^{(\delta,a)}$ and $\boldsymbol{\Gamma}^{(\delta,a)}$ but this was not done in this work.

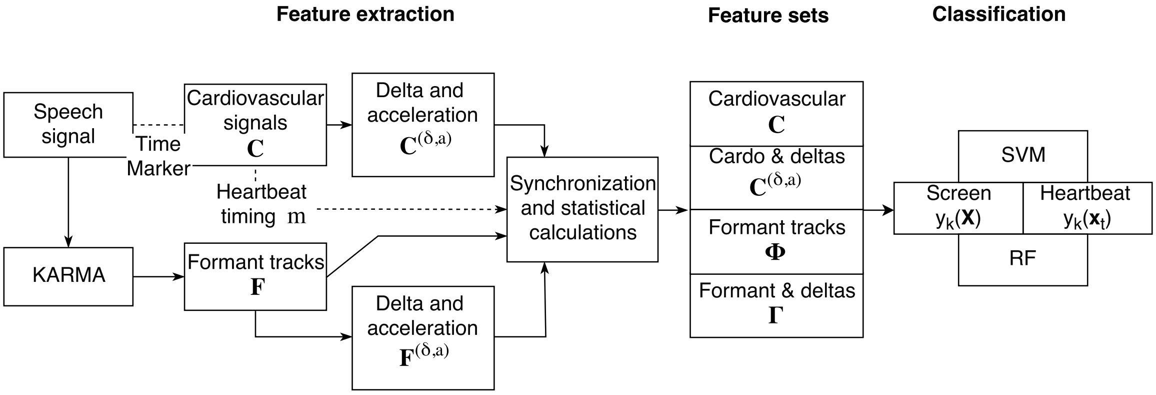


Figure 6.2: Overview of the feature extraction, synchronization and classification process applied in Study III.

6.2 Results

The methods proposed for the combined psychophysiological signals were tested on the cognitive stimuli tasks performed by the participants, i.e. Stroop tasks. As in Study I and Study II, the performance of the different psychophysiological signals on their own in trinary classification is summarized as well as in their combined form. The different supervised learning classifiers are compared with the two CE (screen and heartbeat) in separate tables. Evidence is presented of the possible benefit of combining the cardiovascular and formant track feature sets and the results of these presented in the subsequent Section. The results in Section 6.2.2 support the evidence of there being a benefit to combining the two feature sets with the best performance of 15.17±0.79% MCR±SE, introduced with the combined feature sets with the addition of time dynamic context with delta-delta features. The final section presents a more fine grained analysis of the performance of the three Stroop cognitive workload levels. The best performance results are presented in confusion tables outlining the relationship between the classes for screen and heartbeat classification.

6.2.1 Feature set performance

The classification results for the separate cardiovascular $\mathbf{C}$ and formant track $\mathbf{\Phi}$ feature sets are presented in Table 6.1 for the sequence of feature vectors spanning a single screen $\mathbf{X}$. Similarly, Table 6.2 depicts the results for separate feature sets ($\boldsymbol{\phi}_n$ and $\mathbf{c}_l$) with the classification scheme of heartbeat $\mathbf{x}_n$ classification. The tables compare the performance of the two classifiers SVM and RF and the benefit of adding the dynamic context of the signals $\mathbf{C}^{(\delta,a)}$ and $\mathbf{\Gamma}$ for Table 6.1 and γ_n and $\mathbf{c}_l^{(\delta,a)}$ for Table 6.2. The results are presented as the average MCR for all $P = 97$ participants and the standard error.

In comparing the formant tracks and cardiovascular signals, both Table 6.1 and Table 6.2 show that the cardiovascular features greatly outperform the formant tracks and also that the formant tracks benefit the most from adding time dynamic information to the feature set. These results are consistent with respect to classifier performance, but in terms of comparing the two, RF outperforms SVM.

Table 6.1: Average MCR±SE [%] for all participants with cardiovascular $\mathbf{C}$ and formant track $\mathbf{\Phi}$ feature sets and their corresponding time dynamic $\mathbf{C}^{(\delta,a)}$ and $\mathbf{\Gamma}$ feature sets showing the results for the SVM and RF classifiers with the screen $\mathbf{X}$ classification scheme.

Feature **sets**	**SVM Screen** $\mathbf{y}_{SVM}(\mathbf{X})$	**RF Screen** $\mathbf{y}_{RF}(\mathbf{X})$
$\mathbf{\Phi}$	55.23±1.10	42.07±1.09
$\mathbf{\Gamma}$	46.98±1.11	41.38±1.09
$\mathbf{C}$	26.61±0.98	22.97±0.93
$\mathbf{C}^{(\delta,a)}$	25.97±0.97	23.32±0.94

Table 6.2: Average MCR±SE [%] for all participants with cardiovascular $\mathbf{c}_l$ and formant track $\boldsymbol{\phi}_n$ feature sets and their corresponding time dynamic $\mathbf{c}_l^{(\delta,a)}$ and $\boldsymbol{\gamma}_n$ feature sets showing the results for the SVM and RF classifiers with the heartbeat $\mathbf{x}_n$ classification scheme.

Feature **sets**	**SVM Beat** $\mathbf{y}_{SVM}(\mathbf{x}_n)$	**RF Beat** $\mathbf{y}_{RF}(\mathbf{x}_n)$
$\boldsymbol{\phi}_n$	61.75±0.18	53.50±0.18
$\boldsymbol{\gamma}_n$	57.88±0.18	52.58±0.18
$\mathbf{c}_l$	36.62±0.18	37.57±0.18
$\mathbf{c}_l^{(\delta,a)}$	36.17±0.17	37.48±0.18

6.2.2 Combined signal performance

The results for the combined feature sets are presented in the same manner as the results feature sets in Table 6.1 and Table 6.2. Here Table 6.3 and Table 6.4 respectively state the results of average MCR±SE for combinations of concatenated feature sets, with and without delta δ and acceleration a coefficients.

Table 6.3: Average MCR±SE [%] for all participants with different combinations of cardiovascular $\mathbf{C}$ and formant track $\mathbf{\Phi}$ feature sets and their corresponding time dynamic $\mathbf{C}^{(\delta,a)}$ and $\mathbf{\Gamma}$ feature sets showing the results for the SVM and RF classifiers with the screen $\mathbf{X}$ classification scheme.

Feature **sets**	**SVM Screen** $\mathbf{y}_{SVM}(\mathbf{X})$	**RF Screen** $\mathbf{y}_{RF}(\mathbf{X})$
$\mathbf{\Phi}$ & $\mathbf{C}$	17.77±0.85	18.41±0.86
$\mathbf{\Phi}$ & $\mathbf{C}^{(\delta,a)}$	16.99±0.83	19.29±0.87
$\mathbf{\Gamma}$ & $\mathbf{C}$	15.66±0.81	19.14±0.87
$\mathbf{\Gamma}$ & $\mathbf{C}^{(\delta,a)}$	$\mathbf{15.17 \pm 0.79}$	19.29±0.87

In contrast to the results in previous section, the SVM classifier outperforms the RF in all cases and classification with RF seems even to deliver worse results with each addition of feature sets. The best result for the RF of 18.41 ± 0.86% (Table 6.3) is

Table 6.4: Average MCR±SE [%] for all participants with different combinations of cardiovascular $\mathbf{c}_l$ and formant track $\boldsymbol{\phi}_n$ feature sets and their corresponding time dynamic $\mathbf{c}_l^{(\delta,a)}$ and $\boldsymbol{\gamma}_n$ feature sets showing the results for the SVM and RF classifiers with the heartbeat $\mathbf{x}_n$ classification scheme.

Feature **sets**	**SVM Beat** $\mathbf{y}_{SVM}(\mathbf{x}_n)$	**RF Beat** $\mathbf{y}_{RF}(\mathbf{x}_n)$
$\boldsymbol{\phi}_n$ & $\mathbf{c}_n$	32.57±0.17	34.63±0.17
$\boldsymbol{\phi}_n$ & $\mathbf{c}_l^{(\delta,a)}$	32.43±0.17	34.43±0.17
$\boldsymbol{\gamma}_n$ & $\mathbf{c}_n$	32.16±0.17	35.76±0.17
$\boldsymbol{\gamma}_n$ & $\mathbf{c}_l^{(\delta,a)}$	32.14±0.17	34.57±0.17

achieved with $\mathbf{C}$ combined with $\mathbf{\Phi}$ without time dynamic information. This suggests that the RF classifier is not as suitable for large combined feature sets as the SVM.

The lowest overall average MCR±SE of $\mathbf{15.17} \pm \mathbf{0.79}\%$, was achieved by combining all feature sets, cardiovascular $\mathbf{C}^{(\delta,a)}$ and formant tracks $\mathbf{\Gamma}$ with the SVM classifier and screen CE.

6.2.3 Cognitive workload level comparison

To gain insight into the feasibility of the Stroop tasks as a cognitive stimuli and the specific experiment configuration introduced in this work a confusion table for the three Stroop levels is presented in Table 6.5. In the table this point is demonstrated with the feature extraction and classifier combination resulting in the lowest MCR; the SVM classifier with the cardiovascular $\mathbf{C}^{(\delta,a)}$ and formant tracks $\mathbf{\Gamma}$ and time dynamic combined feature set.

The numbers are obtained by summing up the number of screens (first table) and number of heartbeats (second table) for all participants giving the total number of CE classified to each Stroop level. The tables also depict the average MCR [%] and the average MTR [%] for this feature set combination. The total average test set MCR is also given in bold. The errors,indicated by MTR, seem to be most dominant in level 2, which is not surprising for it lacks the distinction of the other two. However, Stroop level 3 seems to be performing slightly better than the other levels with a 9.15% and 25.98% MCR and 14.75% and 27.55% MTR respectively for screen and heartbeat classification.

6.3 Discussion and Conclusions

The methods of classification introduced in this paper are based on individual participants and as a result are not generic. Despite the promising results published by Yin *et al.* [6] with participant independent methods, these individual differences have been a major confound in prior work.

Even though these two psychophysiological signals have not been paired for cognitive workload classification these results indicate that they greatly compliment each other in the task. In addition, the results in Table 6.5 indicate that most of the incorrect classifications lie in adjacent cognitive workload levels e.g. high miss-classified as medium. Such miss-classification may indicate the boundary between adjacent levels

Table 6.5: Confusion tables for the trinary classification of three Stroop levels. The tables give a closer look at the results for the SVM classification for all participants with the combined feature set, delta and acceleration ($\mathbf{C}^{(\delta,a)}$ & $\mathbf{\Gamma}$. The first table adds all classified screens $\mathbf{X}$ and the second table adds up all heartbeats $\mathbf{x}_n$.

	Classified as			
Actual	**L1**	**L2**	**L3**	**MCR [%]**
Stroop L1	**569**	65	45	16.2
Stroop L2	51	**454**	77	21.99
Stroop L3	23	48	**705**	9.15
MTR [%]	11.51	19.93	14.75	**15.17**

	Classified as			
Actual	**L1**	**L2**	**L3**	**MCR [%]**
Stroop L1	**13810**	3941	3353	34.56
Stroop L2	4252	**15516**	4846	36.96
Stroop L3	3252	4316	**21566**	25.98
MTR [%]	35.21	34.73	27.55	**32.01**

may bee too rigid at a generic level and further highlight the need for granularity at the individual level.

In this study, it has been demonstrated the potential of using a combination of psychophysiological measures, the speech- and cardiovascular signals, to both measure and monitor cognitive workload. Based upon a much larger sample than in any previous study and using carefully designed experiments with increasing task complexity, we have shown that this combination can reliably measure and discriminate three levels of cognitive workload at the level of an individual.

Chapter 7

Study IV Cardiovascular analysis with distance from baseline metrics

In Study IV a different approach is taken to the previous feature extraction methods focusing on identifying time profiles through cardiovascular signals. The hypothesis of concern in Study IV has to do with individual reactions to cognitive workload levels during cognitive workload tasks. The methods used so far have not given insight into what is going on in the cardiovascular system during changes in cognitive workload levels. As discussed in Section 1.4 the dynamic reactions of the cardiovascular system to commands from the ANS and its contradictory but complimentary messages make it a taxing effort. Therefore we ask if it is possible to identify *cardiovascular profiles, describing reactions to cognitive workload levels with a distance measure from baseline? If so, can groups of individuals and/or task levels be identified according to these profiles?* The goal is to be able to evaluate the cardiovascular system as a whole hemodynamic entity and thereby gain insight into cardiovascular reactions to different levels of cognitive workload.

A novel approach is introduced that can be used to aid in the exploration of the relationship between cognitive workload and task loads based on human or task taxonomy. This involves retrieving one distance measure from multi-dimensional cardiovascular measures from a baseline resting state to the cognitive workload tasks. By assuming that each predefined window of ten cardiovascular measures are normally distributed, the distance between two Gaussian probability density functions could be measured. From each task nine profile values were retrieved describing distance and timing e.g. maximum, minimum distance values and duration and grouped with cluster analysis.

7.1 Methods

A distance measure of cognitive workload levels, compared to a normal state, was calculated from the ten cardiovascular signals using Bhattacharyya distance measure. By calculating for the distance from the baseline period to a window of a certain size one distance measure comparing two distributions is achieved. These calculations were then applied to each heartbeat in a sliding manner describing the deviation in effort compared to baseline extracted at the time instance of a heartbeat throughout the experiment.

7.1.1 Distance measures from baseline

With the Bhattacharyya distance measure (BDM) [118] the similarity of two probability distributions are given as one distance value. From the cardiovascular measures matrix $\mathbf{C}$ for each participant with total number of measurements $I \times 10$ the baseline period $\mathbf{C}_B$ was identified. The BDM calculations were then performed with $\mathbf{C}_B$ and a segment containing the line vectors

$$\mathbf{C}_n = (\mathbf{c}_{n-\frac{w}{2}}, ..., \mathbf{c}_n, ..., \mathbf{c}_{n+\frac{w}{2}}) \tag{7.1}$$

from the data matrix $\mathbf{C}$, with a sliding window w for heartbeat n. Assuming for the two segments that $p(\mathbf{C}_B) = \mathcal{N}(\mathbf{c}|\mu_B, \Sigma_B)$ where μ_B and Σ_B are the mean vector and covariance matrix for baseline period and that $p(\mathbf{C}_n) = \mathcal{N}(\mathbf{c}|\mu_n, \Sigma_n)$ where μ_n and Σ_n are the mean vector and covariance matrix for the period of size w for heartbeat n. From these a multivariate Bhattacharyya distance was calculated for each heartbeat n using the formula,

$$b_n = \frac{1}{8}(\mu_B - \mu_n)^T(\frac{\Sigma_B + \Sigma_n}{2})^{-1}(\mu_B - \mu_n) + \frac{1}{2}\ln(\frac{\det \Sigma_n}{\sqrt{\det \Sigma_B \det \Sigma_n}}) \tag{7.2}$$

resulting in a distance measure vector $B = [b_1, b_2, ..., b_N]^T$ with a total number of N heartbeats for each participant.

Figure 7.1: Example of the BDM results during an experiment session for one participant. Here the BDM has been calculated from the baseline session in the beginning with the window size of 50 heartbeats. The shaded areas mark the periods of reading and Stroop levels tasks.

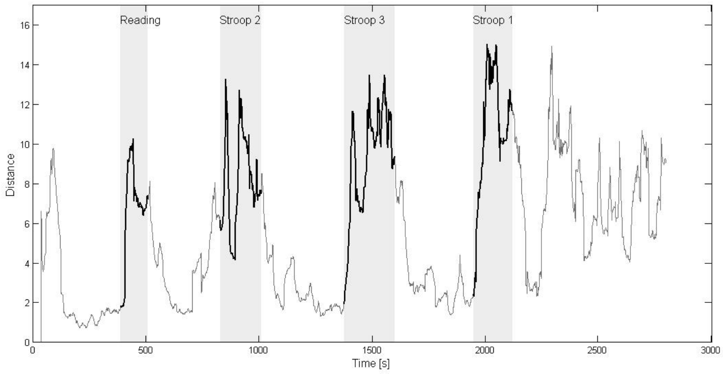


Figure 7.1 shows an example of the BDM throughout the experiment session for one participant. The reactions of the individuals cardiovascular system compared to baseline can be clearly seen during the four task periods (shaded gray). The areas not shaded contain BDM for other sections of the experiment session. The area before the tasks start contain the baseline, the area between task the resting periods and the last period the OSPAN portion of the sessions Figure 3.1 in the methods chapter describes the flow of the experiment session in detail. Extreme increase in distance values can be seen at task onset compared to the resting periods located between tasks, as well as well defined peaks throughout the tasks.

For the purposes of this study a BDM with the window size of 50 heartbeats was chosen. To reduce extreme measures that might have been included in the baseline period due to e.g. interruptions in the beginning of the session, the first 10% of the frames was excluded.

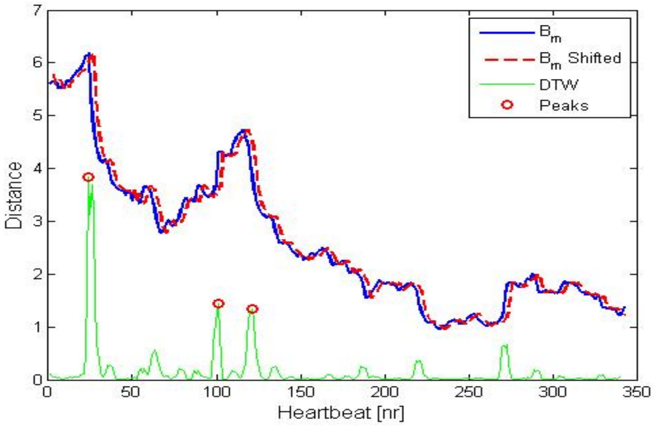
7.1.2 Profile feature extraction

Figure 7.1 shows that the cardiovascular reactivity as measured by the BDM is responsive to tasks. Difficult tasks appear to have more extreme values (further away from the rest condition in baseline) and contain other features such as a second peak and high level of changes. It is therefore appropriate to capture these features numerically. Based on observations, profile features describing timing, distance values, velocity (i.e. change in distance) and acceleration (i.e. change in velocity) of the extreme points of the BDM sequence during the four tasks were retrieved.

The four task periods were labeled $m = 1$ for reading, $m = 2$ for Stroop 1, $m = 3$ for Stroop 2 and $m = 4$ for Stroop 3. In every case reading ($m = 1$) was presented first but the order of the other tasks ($m = 2, 3, 4$) was randomized. The instruction screen was included in the first Stroop task whichever one that happened to be. For each task, the corresponding B distance vector was retrieved and denoted as B_m (see e.g. shaded periods in Fig. 7.1).

In Fig. 7.2 an example of the B_m for one task m is depicted. From these calculations a profile value vector ψ_m containing nine characteristics was retrieved for each task. The first four values representing the distance at start of task $\psi_{1,m}$, maximum distance during task $\psi_{2,m}$, number of heartbeats to the maximum distance $\psi_{3,m}$ and minimum distance $\psi_{4,m}$ were retrieved directly from B_m.

Figure 7.2: The BDM from baseline B_m for one task m along with the same distance measure B_m shifted 3 heartbeats (dashed line). For this particular example three peaks (circles) are located from the DTW cost.



Along with B_m a shifted version of B_m by three heartbeats is plotted as a dotted line in Fig. 7.2. The next three profile values were retrieved by calculating the fitted second order polynomials for segments containing three heartbeats of the two (B_m and shifted B_m) for all heartbeats during the task. From these the maximum $\psi_{5,m}$ and minimum $\psi_{6,m}$ first order polynomials and the maximum $\psi_{7,m}$ second order polynomial were retrieved.

For the profile values $\psi_{8,m}$ and $\psi_{9,m}$ a method used to find alignment between two time-dependent series called dynamic time warping (DTW) was employed [119].

The method uses a local cost function with the goal of finding the optimum warping path between two series with the minimal cost or total distance. DTW values were calculated for the same adjacent three adjacent heartbeat windows comparing B_m and shifted B_m for all heartbeats in during the task. The DTW cost function can be interpreted as the magnitude of difference between the two segments, where high values indicate extreme differences in a six heartbeat period. A peak finding algorithm was used to find the local maximum $\psi_{8,m}$ and second maximum $\psi_{9,m}$ peaks. In Fig. 7.2 the DTW results (green line) from the distance measure to the shifted distance measure is displayed along with the BDM and the peaks (red circle) located from the DTW calculations. From these profile values the corresponding log feature vectors then become,

$$V_m = \log([\psi_{1,m}, \psi_{2,m}, ...\psi_{9,m}])^T \tag{7.3}$$

for task m.

Each of the 96 participants produces four profile vectors to the profile data matrix $\mathbf{V}$ where each line is the profile values of one task $m = (1, 2, 3, 4)$ for one participant $p = (1, ..., 96)$.

7.1.3 Clustering

Two clustering algorithms were used evaluation of the profile value data matrix V_m described in more detail in Section 1.5.2.5 for the k-means and the GMM in Section 1.5.2.4.

To evaluate tightness and separation of clusters a method called silhouettes, introduced by Rousseeuw [120] is used. Each instance from the profile feature vector V_m is represented with a silhouette value from -1 to $+1$ indicating its fit to its assigned cluster. Profile feature vectors with a silhouette value of $+1$ indicate that these features are very distant from neighboring clusters and 0 indicate features that are not distinctly in one cluster or another. A negative -1 value, however, indicates that the features are probably assigned to the wrong cluster. For evaluation of the number of clusters and the fit of these clusters the average silhouette value υ for clusters is depicted in the results.

7.2 Results

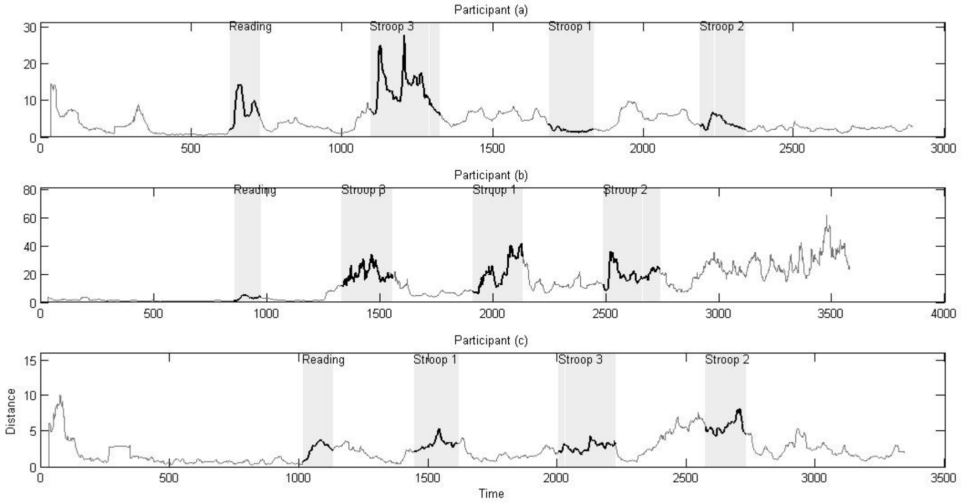
The results of Study IV are represented through two methods, first the the apparent variation in BDM will be interpreted. The sessions for three individuals are interpreted and the possible reasons behind different distance measure profiles discussed. Secondly, results for the clustering trials of the data matrix $\mathbf{V}$ will be are presented for two levels of clustering, individual- and task based.

7.2.1 Bhattacharyya distance measures

Graphical evaluation of the BDM illustrates quite a few things when comparing both individuals as well as tasks. As evident from Fig. 7.3 both extreme increases as well as more level reactions in distance value from baseline can be detected during tasks as opposed to other parts of the sessions.

From the figure the different reactions during the session become apparent. Not only can different profiles bee seen between individuals through variations in the quan-

Figure 7.3: BDM from the baseline during sessions for three participants a, b and c. Different reactions to cognitive workload from baseline are apparent for individuals and tasks when comparing these three participants.



tity of distance from baseline but also between tasks (shaded gray). Different profiles of BDM quantities can be detected within tasks as well as between tasks according to the difficulty of the tasks (Stroop levels or reading). Within some tasks, there seems to be more than one climax period where the peak distance quantity is reached quickly and then again later on during the task. Participant (a) e.g. has two distinct reaction peaks during Stroop 3, indicating that this level of difficulty caused extreme reactions for this individual. However, the distance quantities during Stroop 1 and 2 are much lower and even lower than the reactions to the reading task. Whether the reasons for this are that this participant has overcome his initial anxiety after the Stroop 3 task, has given up and relaxed on the effort for the remainder of the tasks or simply found Stroop 1 and 2 so much easier remains unknown.

Perhaps the order of the Stroop levels influenced the reaction profile for Participant (a), however this seems not to be the case for Participant (b). Although receiving the Stroop levels in the same order and showing high distance values, this individual displays effort throughout all the Stroop levels and comparatively very low during reading. Participant (b) also does not reach many distinct climax peaks but displays rather more erratic reactions during tasks.

In contrast, Participant (c) has a lower distance profile compared to the other two throughout the whole session. There seems to be, however, a steady increase in the quantities of distance from baseline during the resting period between Stroop 3 and Stroop 2. As a result the Stroop 2 distance starts relatively high and becomes the highest for this individual. The cause might be that the anticipation for the next task is building steadily because of the difficulty of the previous one or perhaps fatigue becomes an issue this late in the session. These possible explanations need, however, further controlled experimentation in order to be answered successfully.

7.2.2 Individual level clustering

With the individual level clustering the profile data matrix **V** is clustered as described earlier, i.e. each line is the profile values of one task $m = (1, 2, 3, 4)$ for one participant $p = (1, ..., 96)$. The individual level clustering results are depicted in Fig. 7.4 for k-means clustering and Fig. 7.5 for GMM clustering with a silhouette plot and the average silhouette value υ for each case.

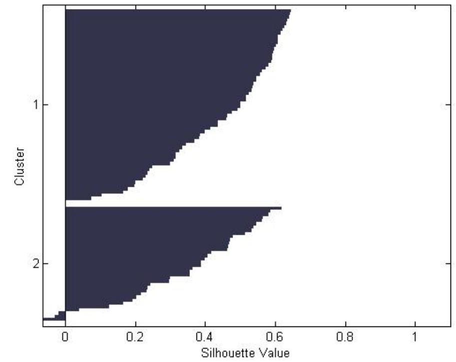


Figure 7.4: Silhouette for individual level clustering depicting the fit of two clusters with average distance of $\upsilon = 0.41041$ with k-means clustering.

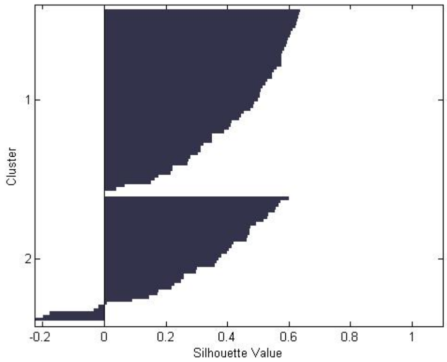


Figure 7.5: Silhouette for individual level clustering depicting the fit of two clusters with average distance of $\upsilon = 0.38928$ with GMM clustering.

As the average silhouette values indicate two groups of individual profiles can be detected with this method with k-means $\upsilon = 0.41041$ compared to GMM $\upsilon = 0.38928$. Comparing the two clustering methods for individual based clustering, the k-means algorithm achieves slightly better average silhouette value υ than the GMM.

7.2.3 Task level clustering

For the task level clustering the profile data matrix **V** dimensions are set up to form a profile value vector for all tasks regardless of participants. In the case of task level clustering three clusters achieved the best results for the GMM algorithm. However the k-means algorithms reached the best result for four clusters. Fig. 7.6 and 7.7 shows the silhouettes for the optimum number of clusters for the two clustering methods.

To take a closer look at the clustering according to tasks a similar technique to the confusion table for classification is introduced in Tables 7.1 and 7.2. These tables depict the number of tasks that are clustered together as opposed to the actual classes to classification results in confusion tables.

Comparing the Tables 7.1 and 7.2 the reasons for the four cluster setup outperforming the three cluster setup becomes clear for k-means clustering. The profile values grouped together in cluster 2 (C2 in Table 7.2) seem to have had decremental effect on the average silhouette value when clustered with the other three clusters. The GMM clustering algorithm, however, is not able to detect this cluster and therefore is not able to achieve better results for the four cluster setup.

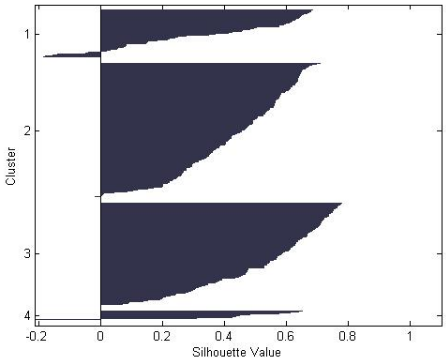


Figure 7.6: K-means silhouette depicting the optimum fit of four clusters achieving the average silhouette value $\upsilon = 0.45365$.

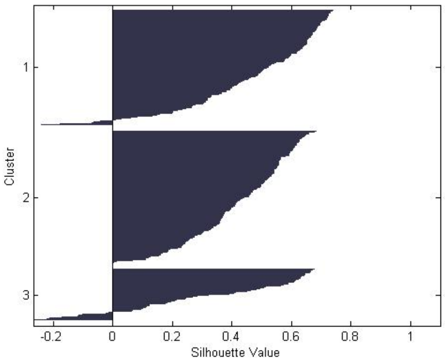


Figure 7.7: GMM silhouette depicting the optimum fit of three clusters achieving the average silhouette value $\upsilon = 0.42739$.

Table 7.1: K-means clustering performance according to tasks with three clusters. The average silhouette value $\upsilon = 0.44049$.

	C1	**C2**	**C3**
Reading	49	9	38
Stroop 1	38	11	47
Stroop 2	36	18	42
Stroop 3	16	26	54

Table 7.2: K-means clustering performance according to tasks with four clusters. The average silhouette value $\upsilon = 0.45365$.

	C1	**C2**	**C3**	**C4**
Reading	9	0	38	49
Stroop 1	11	6	45	34
Stroop 2	18	6	40	32
Stroop 3	25	1	52	18

7.3 Discussion and Conclusions

By exploring the distance measures graphically different hemodynamic reactions to cognitive workload can be seen, not only between individuals but also between tasks. Thereby establishing that cardiovascular profiles can be identified through distance measures from baseline. It seems, however that numerical interpretation of profile values from these distance measures is lacing with the methods used. Even though the BMD profile values features can be extracted and clustered into two on individual based clustering and four on task based clustering. This can not be considered good results compared to e.g. the classification results presented in Study II and III of this work.

Through these results the differences in individual reactions to cognitive workload become apparent and the difficulties of joint overall mathematical interpretation highlighted. This further establishes the cause of the leading confound in the cognitive workload level detection research; the problem of individual differences.

Part III

Discussion and Conclusions

Chapter 8

Discussion

The most prominent results from each of the four studies constructing this work are summarized in Table 8.1. From the table, a steady improvement pattern can be seen from study to study lowering the MCR to the optimum feature and classification combination to 15.17% MCR for screen based trinary classification scheme.

Table 8.1: Main findings of the four studies along with the average MCR [%] for the best achieving combination of feature sets and classification methods.

	Optimum method	**Average** MCR[%]
Study I	Weighted vocal tract - and voice source features fused at the output level with SVM classification.	32.5
Study II	Temporal cardiovascular features with screen based SVM classification.	20.44
Study III	Heartbeat synchronized temporal feature sets with screen based SVM classification.	15.17
Study IV	Individual cardiovascular profiles Bhattacharyya distance from baseline.	-

The chosen psychophysiological signals play a crucial role for the performance of a cognitive workload classification model. Their sensitivity to changes in cognitive workload are a crucial factor and their intrusiveness during retrieval has to be minimal for real-world purposes. Combining two or more signals is an excellent strategy for the models performance improvement. Much effort has been put into researching different combinations of psychophysiological measures such as with electrical brain waves and/or cardiovascular signals, especially in the context of aviation. However, the two opportune measures (cardiovascular- and speech signals) investigated in this work introduce a combination not researched before.

By employing the delta time variation method, known from speech processing, a different approach is introduced for cardiovascular variability evaluations. With this method the amount of data behind each measurement is reduced to 2 adjacent heartbeats as opposed to traditional methods such as e.g. [3], [121]. This, as the results indicate, proved to be a step towards incorporating the influence of the short-term reactions of the ANS on the cardiovascular system.

Although an average MCR of 15.17% may seem to be a barrier, the results in Study III indicate that most of the incorrect classifications lie in adjacent cognitive workload levels e.g. high misclassified as medium. Such misclassification may indicate the boundary between adjacent levels may be too rigid at a generic level and further highlight the need for granularity at the individual level. Even though not achieving ideal results, trinary classification for unknown heartbeats during tasks was attempted and can perhaps along with regression instead of classification move the research field towards granular results.

Another essential part of this work is the introduction of a novel method to describe groups of individuals with respect to their cognitive workload reactions. The aim was to address the question *can individuals reactions to cognitive workload be grouped according to some common characteristics that can be used to explain and/or reduce individual differences?* Perhaps by calculating a distance measure from the individuals baseline and using it as a measure for cognitive workload tasks, as suggested in Study IV, might move us closer to this understanding?

The methods of classification introduced in this work are based on individual participants and as a result are not generic. Despite the promising results published by Yin *et al.* [6] with participant independent methods, these individual differences have been a major confound in prior work. Although achieving high accuracy, Yin *et al.*'s [6], [64] results must be treated with caution. Therefore a participant dependent classification scheme supplemented with pattern recognition algorithms enabling the model to be further adapted to the individual is proposed as the way forward.

The experimental setup and the choices made in the design of the process might be subject to critique. The limited vocabulary of the Stroop task, where the participants uttered the same five words throughout, might seem too limiting for real-world operational scenarios. However, the features extracted are not based on linguistics but rather on the amplitude peaks in certain frequency spectra. This can therefore be the basis for a model intended for aviation personnel, limited to the vocabulary used internationally in air traffic communication terminology. Testing in real-life circumstances was not within the scope of this particular work, but remains a part of the long term research goals. The fact that the same participants were not asked to perform the session more than once for comparison was a decision made during the designing process. The learning effect of the Stroop task between sessions (e.g. [122]) was the main reason for this and therefore one detailed experimental session chosen instead.

The work in Study I was heavily based on the work of the the NICTA [6] groups publications, however an exact replication for comparisons was not attempted. There might be several reasons for the results of the speech feature extraction methods introduced in Study I (see Table 8.1) not quite reaching the results as listed in Table 2.1 for closed set results. They are however, on par or not far behind the open set results listed for formants (32.2% MCR from NICTA compared to 33.92% MCR) and even better than the combined formants (vocal tract) and voice source features (37.3% MCR from NICTA compared to 32.5% MCR with weighted combination of the features in Study I) . In hindsight, a replication of the results of the NICTA methods might have been good, however based on the results from Study I the decision was made to expand our methods by exploring other avenues.

Chapter 9

Conclusions

In safety critical operations, such as air traffic control, the accurate measurement and monitoring of the cognitive workload of operators is a pre-requisite for safety and efficiency. However, the traditional methods of measurement based upon either self-assessed subjective ratings or on performance measures of tasks have their well-documented limitations. The use of psychophysiological measurement is another technique that has failed to gain widespread acceptance in the past due to reliability and logistical issues. However, recent advances in technology and computational abilities provide an opportunity to examine this technique anew.

The research questions set for this work were:

- Can cognitive workload be detected from speech- or cardiovascular signals and if so can these signals be incorporated to move beyond binary distinction?
- Do these two psychophysiological signals compliment each other to improve cognitive workload classification?
- Are there methods that can better help us to understand the complex reactions of the cardiovascular system and thus to better monitor cognitive workload?

Each of these research questions have been addressed throughout the four studies presented in this work. First, that signals from both psychophysiological sources, speech- and cardiovascular, can be used efficiently to detect three cognitive workload levels. Even though cardiovascular signal classification proved to achieve better results, the research in the field of speech signals with respect to cognitive workload is quite less developed and is therefore subject to even more improvements. The combination and synchronization of the signals proved to be beneficial and thereby introducing a combination never researched before in the field of cognitive workload modeling. In response to the final research question, a novel method using a multivariate Bhattacharyya distance method, between baseline and task periods, to capture individuals hemodynamic profile reactions to cognitive workload is introduced. An obvious difference can be detected from these distance measure results, both between individuals and tasks. Even though attempts to interpret profile values from these distance values were not fruitful they demonstrate the potential use of these measures in the future.

In this work, we have demonstrated the potential of using a combination of psychophysiological measures, the speech signal and cardiovascular measures, to both measure and monitor cognitive workload. Based upon a much larger validation sample than in any previous study and using carefully designed experiments with increasing task complexity, we have shown that this combination can reliably measure and discriminate three levels of cognitive workload at the level of an individual. This then provides safety critical organizations with the flexibility to manage the workload of their personnel and thereby enhance safety.

The proposed cognitive workload classification and feature extraction approach provides an important contribution as a starting point for a model that can be customized to the individual. To reach the ultimate goal of a real-world application we believe that these method should be extended with a learning algorithm trained for the individual being monitored. In addition, the contribution of a distance measure based on the Bhattacharyya distance measure provides a novel approach towards further research into understanding and modeling of reactions from the cardiovascular system which are highly relevant in multiple fields.

By integrating the ever increasing improvements in computational capacity of computers and applying psychophysiological measures to monitor cognitive workload we are able to move beyond the traditional approaches of self assessment tests and performance evaluations into a highly automatic real-time monitoring model using psychophysiological measures. The theory is that by investigating within heartbeat reactions controlled by the PNS the finer changes in cognitive workload can be better established. Furthermore, that by investigating the time of onset of the ANS the individual's 'danger zone' can be identified.

It is likely that including more than two well chosen signals could prove to be redundant to the models performance and research should focus on increasing the performance of the algorithm used and understanding of the psychophysiolgical signals. In the task of overcoming the remaining $15 - 20\%$ MCR, focus should be set on developing methods that can be used to account for individual human differences.

Bibliography

[1] M. Meier, M. Borsky, E. H. Magnusdottir, K. R. Johannsdottir, and J. Gudnason, “Vocal tract and voice source features for monitoring cognitive workload”, in *Cognitive Infocommunications (CogInfoCom), 2016 7th IEEE International Conference on*, IEEE, 2016, pp. 000 097–000 102.

[2] E. H. Magnusdottir, M. Borsky, M. Meier, K. Johannsdottir, and J. Gudnason, “Monitoring cognitive workload using vocal tract and voice source features”, *Periodica Polytechnica Electrical Engineering and Computer Science*, May 23, 2017.

[3] M. Malik, “Heart rate variability”, *European Heart Journal*, vol. 17, pp. 354–381, 1996.

[4] G. J. Tortora and M. Nielsen, *Principles of human anatomy, 14th edition.* Wiley, Nov. 16, 2016, 987 pp.

[5] G. F. Wilson and C. A. Russell, “Operator functional state classification using multiple psychophysiological features in an air traffic control task”, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 45, no. 3, pp. 381–389, Sep. 1, 2003.

[6] B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah, “Speech-based cognitive load monitoring system”, in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, IEEE, 2008, pp. 2041–2044.

[7] R. Lysaght and S. Hill, “Operator workload: Comprehensive review and evaluation of operator workload methodologies. .”, *Fort Bliss, Texas, U.S. Army Research Institute for the Behavioural and Social Sciences*, p. 262, 1989.

[8] B. H. Kantowitz and O. Simsek, *Secondary-task measures of driver workload (chapter 2.10, 395-408) in hancock, PA and desmond, PA stress, workload, and fatigue.* Mahwah, NJ: Lawrence Erlbaum Associates, 2001.

[9] A. F. Kramer, “Physiological metrics of mental workload: A review of recent progress”, *Multiple-task performance*, pp. 279–328, 1991.

[10] N. Moray, “Models and measures of mental workload”, in *Mental Workload*, Springer, 1979, pp. 13–21.

[11] G. R. J. Hockey, *Operator functional state: The assessment and prediction of human performance degradation in complex tasks.* IOS Press, 2003, vol. 355.

[12] D. Kahneman, *Attention and effort.* Prentice-Hall Englewood Cliffs, NJ, 1973, vol. 1063.

[13] N. Lavie, “Distracted and confused?: Selective attention under load”, *Trends in Cognitive Sciences*, vol. 9, no. 2, pp. 75–82, Feb. 1, 2005.

[14] N. Dunn and A. Williamson, “Driving monotonous routes in a train simulator: The effect of task demand on driving performance and subjective experience”, *Ergonomics*, vol. 55, no. 9, pp. 997–1008, 2012.

[15] R. L. Boring, C. D. Griffith, and J. C. Joe, “The measure of human error: Direct and indirect performance shaping factors”, in *Human Factors and Power Plants and HPRCT 13th Annual Meeting, 2007 IEEE 8th*, IEEE, 2007, pp. 170–176.

[16] J. H. Hansen, C. Swail, A. J. South, R. K. Moore, H. Steeneken, E. J. Cupples, T. Anderson, C. R. Vloeberghs, I. Trancoso, and P. Verlinde, “The impact of speech under ‘stress’ on military speech technology”, *NATO Project Report*, 2000.

[17] G. B. Reid and T. E. Nygren, “The subjective workload assessment technique: A scaling procedure for measuring mental workload”, in *Advances in psychology*, vol. 52, Elsevier, 1988, pp. 185–218.

[18] S. G. Hart and L. E. Staveland, “Development of NASA-TLX (task load index): Results of empirical and theoretical research”, *Advances in psychology*, vol. 52, pp. 139–183, 1988.

[19] B. Hilburn and P. Jorna, “Workload and air traffic control”, *Stress, workload, and fatigue. Mahwah, NJ: L. Erlbaum*, 2001.

[20] J. A. Bargh and E. Morsella, “The unconscious mind”, *Perspectives on psychological science*, vol. 3, no. 1, pp. 73–79, 2008.

[21] R. E. Nisbett and T. D. Wilson, “Telling more than we can know: Verbal reports on mental processes.”, *Psychological review*, vol. 84, no. 3, p. 231, 1977.

[22] A. P. Martins, “A review of important cognitive concepts in aviation”, *Aviation*, vol. 20, no. 2, pp. 65–84, 2016.

[23] J. B. Brookings, G. F. Wilson, and C. R. Swain, “Psychophysiological responses to changes in workload during simulated air traffic control”, *Biological Psychology*, Psychophysiology of Workload, vol. 42, no. 3, pp. 361–377, Feb. 5, 1996.

[24] W. B. Verwey and H. A. Veltman, “Detecting short periods of elevated workload: A comparison of nine workload assessment techniques.”, *Journal of experimental psychology: Applied*, vol. 2, no. 3, p. 270, 1996.

[25] J. D’Arcy and P. Della Rocco, “Air traffic control specialist decision making and strategic planning—a field survey (DOT/FAA/CT-TN01/05)”, *Atlantic City International Airport,(NJ: DOT/FAA William J. Hughes Technical Center*, 2001.

[26] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, “Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness”, *Neuroscience & Biobehavioral Reviews*, Applied Neuroscience: Models, methods, theories, reviews. A Society of Applied Neuroscience (SAN) special issue. Vol. 44, pp. 58–75, Jul. 2014.

[27] A.-M. Brouwer, M. A. Hogervorst, J. B. Van Erp, T. Heffelaar, P. H. Zimmerman, and R. Oostenveld, “Estimating workload using EEG spectral power and ERPs in the n-back task”, *Journal of neural engineering*, vol. 9, no. 4, p. 045 008, 2012.

[28] A. Stuiver, D. De Waard, K. A. Brookhuis, C. Dijksterhuis, B. Lewis-Evans, and L. J. M. Mulder, "Short-term cardiovascular responses to changing task demands", *International Journal of Psychophysiology*, vol. 85, no. 2, pp. 153–160, 2012.

[29] F. Melgani and Y. Bazi, "Classification of electrocardiogram signals with support vector machines and particle swarm optimization", *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 5, pp. 667–677, Sep. 2008.

[30] J. Taelman, S. Vandeput, E. Vlemincx, A. Spaepen, and S. Van Huffel, "Instantaneous changes in heart rate regulation due to mental load in simulated office work", *European journal of applied physiology*, vol. 111, no. 7, pp. 1497–1505, 2011.

[31] J. Vogt, T. Hagemann, and M. Kastner, "The impact of workload on heart rate and blood pressure in en-route and tower air traffic control", *Journal of psychophysiology*, vol. 20, no. 4, pp. 297–314, 2006.

[32] G. F. Wilson, "An analysis of mental workload in pilots during flight using multiple psychophysiological measures", *International Journal of Aviation Psychology*, vol. 12, no. 1, pp. 3–18, Jan. 2002.

[33] R. Lew, "Assessing cognitive workload from multiple physiological measures using wavelets and machine learning", PhD thesis, University of Idaho, 2014.

[34] B. Mehler, B. Reimer, and M. Zec, "Defining workload in the context of driver state detection and HMI evaluation", in *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, ACM, 2012, pp. 187–191.

[35] M. W. Scerbo, "Stress, workload, and boredom in vigilance: A problem and an answer.", *Stress, workload, and fatigue*, 2001.

[36] A. C. Dirican and M. Göktürk, "Psychophysiological measures of human cognitive states applied in human computer interaction", *Procedia Computer Science*, vol. 3, pp. 1361–1367, 2011.

[37] B. Cowley, M. Filetti, K. Lukander, J. Torniainen, A. Henelius, L. Ahonen, O. Barral, I. Kosunen, T. Valtonen, M. Huotilainen, N. Ravaja, and G. Jacucci, "The psychophysiology primer: A guide to methods and a broad review with a focus on human–computer interaction", *Foundations and Trends® in Human–Computer Interaction*, vol. 9, no. 3, pp. 151–308, Nov. 3, 2016.

[38] A. H. ROSCOE, "Heart rate as a psychophysiological measure for in-flight workload assessment", *Ergonomics*, vol. 36, no. 9, pp. 1055–1062, Sep. 1, 1993.

[39] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals & systems (2nd ed.)* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1996.

[40] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition", *ARXIV:1412.5567 [cs]*, Dec. 17, 2014. arXiv: `1412.5567`.

[41] T. F. Yap, J. Epps, E. Ambikairajah, and E. H. C. C. (Member), "Formant frequencies under cognitive load: Effects and classification", *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, p. 219 253, Dec. 1, 2011.

[42] M. A. Anusuya and S. K. Katti, “Front end analysis of speech recognition: A review”, *International Journal of Speech Technology*, vol. 14, no. 2, pp. 99–145, Jun. 1, 2011.

[43] H. Beigi, *Fundamentals of speaker recognition.* Springer Customer Service Center Gmbh, May 1, 2016, 1006 pp.

[44] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals.* New York: Institute of Electrical and Electronics Engineers, 2000, 908 pp.

[45] R. E. Kalman, “A new approach to linear filtering and prediction problems”, *Journal of basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

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

[47] Y.-R. Chien, D. D. Mehta, J. Gudnason, M. Zanartu, and T. F. Quatieri, “Evaluation of glottal inverse filtering algorithms using a physiologically based articulatory speech synthesizer”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1718–1730, 2017.

[48] J. Gudnason and M. Brookes, “Voice source cepstrum coefficients for speaker identification”, in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2008, pp. 4821–4824.

[49] F. Shaffer, R. McCraty, and C. L. Zerr, “A healthy heart is not a metronome: An integrative review of the heart’s anatomy and heart rate variability”, *Frontiers in psychology*, vol. 5, p. 1040, 2014.

[50] G. D. Clifford, “Signal processing methods for heart rate variability”, PhD thesis, University of Oxford Oxford, 2002.

[51] C. Bishop, *Pattern recognition and machine learning*, ser. Information Science and Statistics. New York: Springer-Verlag, 2006.

[52] S. J. Russell, P. Norvig, J. F. Canny, J. M. Malik, and D. D. Edwards, *Artificial intelligence: A modern approach*, 9. Prentice hall Upper Saddle River, 2003, vol. 2.

[53] D. M. Powers, “Evaluation: From precision, recall and f-measure to ROC, informedness, markedness and correlation”, 2011.

[54] D. G. Altman and J. M. Bland, “Standard deviations and standard errors”, *BMJ : British Medical Journal*, vol. 331, no. 7521, p. 903,

[55] P. Besson, E. Dousset, C. Bourdin, L. Bringoux, T. Marqueste, D. R. Mestre, and J. L. Vercher, “Bayesian network classifiers inferring workload from physiological features: Compared performance”, in *2012 IEEE Intelligent Vehicles Symposium*, Jun. 2012, pp. 282–287.

[56] C. Elkin, S. Nittala, and V. Devabhaktuni, “Fundamental cognitive workload assessment: A machine learning comparative approach”, in *Advances in Neuroergonomics and Cognitive Engineering*, ser. Advances in Intelligent Systems and Computing, Springer, Cham, Jul. 17, 2017, pp. 275–284.

[57] A. Fong, C. Sibley, A. Cole, C. Baldwin, and J. Coyne, "A comparison of artificial neural networks, logistic regressions, and classification trees for modeling mental workload in real-time", in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 54, SAGE, 2010, pp. 1709–1712.

[58] N. Cristianini and B. Scholkopf, "Support vector machines and kernel methods: The new generation of learning machines", *Ai Magazine*, vol. 23, no. 3, p. 31, 2002.

[59] L. Breiman, "Random forests", *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[60] T. G. Dietterich, "Ensemble learning", *The handbook of brain theory and neural networks*, vol. 2, pp. 110–125, 2002.

[61] N. Desai, K. Dhameliya, and V. Desai, "Feature extraction and classification techniques for speech recognition: A review", *International Journal of Emerging Technology and Advanced Engineering*, vol. 13, no. 12, pp. 367–371, 2013.

[62] T. F. Yap, J. Epps, E. H. C. Choi, and E. Ambikairajah, "Glottal features for speech-based cognitive load classification", in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2010, pp. 5234–5237.

[63] T. F. Yap, J. Epps, E. Ambikairajah, and E. H. Choi, "Voice source features for cognitive load classification", in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, IEEE, 2011, pp. 5700–5703.

[64] 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, Apr. 2011.

[65] P. N. Le, J. Epps, E. H. Choi, and E. Ambikairajah, "A study of voice source and vocal tract filter based features in cognitive load classification", IEEE, Aug. 2010, pp. 4516–4519.

[66] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 computational paralinguistics challenge: Cognitive & physical load.", in *INTERSPEECH*, 2014, pp. 427–431.

[67] M. Van Segbroeck, R. Travadi, C. Vaz, J. Kim, M. P. Black, A. Potamianos, and S. S. Narayanan, "Classification of cognitive load from speech using an i-vector framework.", in *INTERSPEECH*, 2014, pp. 751–755.

[68] J. R. Williamson, D. W. Bliss, D. W. Browne, and J. T. Narayanan, "Seizure prediction using EEG spatiotemporal correlation structure", *Epilepsy & Behavior*, vol. 25, no. 2, pp. 230–238, Oct. 2012.

[69] R. M. Rose and L. F. Fogg, "Definition of a responder: Analysis of behavioral, cardiovascular, and endocrine responses to varied workload in air traffic controllers.", *Psychosomatic medicine*, vol. 55, no. 4, pp. 325–338, 1993.

[70] P. G. A. M. JORNA, "Heart rate and workload variations in actual and simulated flight", *Ergonomics*, vol. 36, no. 9, pp. 1043–1054, Sep. 1, 1993.

[71] D. B. Kaber, C. M. Perry, N. Segall, and M. A. Sheik-Nainar, "Workload state classification with automation during simulated air traffic control", *The International Journal of Aviation Psychology*, vol. 17, no. 4, pp. 371–390, 2007.

[72] E. A. Byrne and R. Parasuraman, “Psychophysiology and adaptive automation”, *Biological psychology*, vol. 42, no. 3, pp. 249–268, 1996.

[73] A. M. Van Roon, L. J. Mulder, M. Althaus, and G. Mulder, “Introducing a baroreflex model for studying cardiovascular effects of mental workload”, *Psychophysiology*, vol. 41, no. 6, pp. 961–981, Nov. 1, 2004.

[74] R. W. Backs, “Going beyond heart rate: Autonomic space and cardiovascular assessment of mental workload”, *The International Journal of Aviation Psychology*, vol. 5, no. 1, pp. 25–48, Jan. 1, 1995.

[75] K. R. Johannsdottir, E. H. Magnusdottir, S. Sigurjonsdóttir, and J. Gudnason, “Cardiovascular monitoring of cognitive workload: Exploring the role of individuals’ working memory capacity.”, *Biological psychology*, 2017.

[76] A. J. Elliot, V. Payen, J. Brisswalter, F. Cury, and J. F. Thayer, “A subtle threat cue, heart rate variability, and cognitive performance”, *Psychophysiology*, vol. 48, no. 10, pp. 1340–1345, 2011.

[77] S. P. Muthukrishnan, J. P. Gurja, and R. Sharma, “Does heart rate variability predict human cognitive performance at higher memory loads?”, *Indian J Physiol Pharmacol*, vol. 61, no. 1, pp. 14–22, 2017.

[78] F. Chen, J. Zhou, Y. Wang, K. Yu, S. Z. Arshad, A. Khawaji, and D. Conway, *Robust multimodal cognitive load measurement*, ser. Human–Computer Interaction Series. Springer International Publishing, 2016.

[79] L. J. Trejo, N. J. McDonald, R. Matthews, and B. Z. Allison, “Experimental design and testing of a multimodal cognitive overload classifier”, *Foundations of Augmented Cognition*, pp. 13–22, 2007.

[80] G. F. Wilson, J. Estepp, and I. Davis, “A comparison of performance and psychophysiological classification of complex task performance”, in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 53, SAGE, 2009, pp. 141–145.

[81] M. Stikic, C. Berka, D. J. Levendowski, R. F. Rubio, V. Tan, S. Korszen, D. Barba, and D. Wurzer, “Modeling temporal sequences of cognitive state changes based on a combination of EEG-engagement, EEG-workload, and heart rate metrics”, *Frontiers in Neuroscience*, vol. 8, Nov. 5, 2014.

[82] H. Zhang, Y. Zhu, J. Maniyeri, and C. Guan, “Detection of variations in cognitive workload using multi-modality physiological sensors and a large margin unbiased regression machine”, in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2014, pp. 2985–2988.

[83] K. Ryu and R. Myung, “Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic”, *International Journal of Industrial Ergonomics*, vol. 35, no. 11, pp. 991–1009, 2005.

[84] M. De Rivecourt, M. N. Kuperus, W. J. Post, and L. J. M. Mulder, “Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight”, *Ergonomics*, vol. 51, no. 9, pp. 1295–1319, 2008.

[85] M. Grassmann, E. Vlemincx, A. von Leupoldt, and O. Van den Bergh, “Individual differences in cardiorespiratory measures of mental workload: An investigation of negative affectivity and cognitive avoidant coping in pilot candidates”, *Applied Ergonomics*, vol. 59, pp. 274–282, 2017.

[86] J.-C. Junqua and J.-P. Haton, “Speaker variability and specificity”, in *Robustness in Automatic Speech Recognition*, ser. The Kluwer International Series in Engineering and Computer Science, Springer, Boston, MA, 1996, pp. 127–153.

[87] G. Matthews, L. Reinerman-Jones, J. Abich, and A. Kustubayeva, “Metrics for individual differences in EEG response to cognitive workload: Optimizing performance prediction”, *Personality and Individual Differences*, Robert Stelmack: Differential Psychophysiology, vol. 118, pp. 22–28, Nov. 1, 2017.

[88] E. H. Magnusdottir, K. R. Johannsdottir, C. Bean, B. Olafsson, and J. Gudnason, “Cognitive workload classification using cardiovascular measures and dynamic features”, in *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, Debrecen: IEEE, Sep. 2017, pp. 000 351–000 356.

[89] M. H. L. Hecker, K. N. Stevens, G. von Bismarck, and C. E. Williams, “Manifestations of task-induced stress in the acoustic speech signal”, *The Journal of the Acoustical Society of America*, vol. 44, no. 4, pp. 993–1001, Oct. 1, 1968.

[90] A. L. Hansen, B. H. Johnsen, and J. F. Thayer, “Vagal influence on working memory and attention”, *International Journal of Psychophysiology*, vol. 48, no. 3, pp. 263–274, 2003.

[91] G. G. Berntson, J. T. Cacioppo, P. F. Binkley, B. N. Uchino, K. S. Quigley, and A. Fieldstone, “Autonomic cardiac control. III. psychological stress and cardiac response in autonomic space as revealed by pharmacological blockades”, *Psychophysiology*, vol. 31, no. 6, pp. 599–608, 1994.

[92] M. E. Gregg, T. A. Matyas, and J. E. James, “A new model of individual differences in hemodynamic profile and blood pressure reactivity”, *Psychophysiology*, vol. 39, no. 1, pp. 64–72, 2002.

[93] J. E. James, M. E. D. Gregg, T. A. Matyas, B. M. Hughes, and S. Howard, “Stress reactivity and the hemodynamic profile–compensation deficit (HP–CD) model of blood pressure regulation”, *Biological psychology*, vol. 90, no. 2, pp. 161–170, 2012.

[94] J. R. Stroop, “Studies of interference in serial verbal reactions.”, *Journal of experimental psychology*, vol. 18, no. 6, p. 643, 1935.

[95] 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 user’s guide”, *Psychonomic bulletin & review*, vol. 12, no. 5, pp. 769–786, 2005.

[96] Finapres Medical Systems. (). Finometer® PRO, [Online]. Available: `http://www.finapres.com/Products/Finometer-PRO` (visited on 10/17/2018).

[97] I. Guelen, B. E. Westerhof, G. L. van der Sar, G. A. van Montfrans, F. Kiemeneij, K. H. Wesseling, and W. J. Bos, “Finometer, finger pressure measurements with the possibility to reconstruct brachial pressure”, *Blood pressure monitoring*, vol. 8, no. 1, pp. 27–30, 2003.

[98] *BeatScope 1.1*, Arnheim, The Netherlands, May 2002.

[99] N. Kolev and M. Zimpfer, “Left ventricular ejection time and end-systolic pressure revisited”, *Anesthesia & Analgesia*, vol. 81, no. 4, p. 889, Oct. 1995.

[100] J. Sundberg, “Vocal intensity in speakers and singers”, *N CVS Status and Progress Report*, p. 17, 1992.

[101] D. G. Childers and C. K. Lee, “Vocal quality factors: Analysis, synthesis, and perception”, *The Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, Nov. 1, 1991.

[102] P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound”, in *Proceedings of the institute of phonetic sciences*, vol. 17, Amsterdam, 1993, pp. 97–110.

[103] Y. D. Heman-Ackah, R. T. Sataloff, G. Laureyns, D. Lurie, D. D. Michael, R. Heuer, A. Rubin, R. Eller, *et al.*, “Quantifying the cepstral peak prominence, a measure of dysphonia”, *Journal of Voice*, vol. 28, no. 6, pp. 783–788, 2014.

[104] P. Alku, “Glottal wave analysis with pitch synchronous iterative adaptive filtering”, *J*, vol. 11, pp. 109–118, 1992.

[105] D. L. Thomson and R. Chengalvarayan, “Use of periodicity and jitter as speech recognition features”, in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1, IEEE, 1998, pp. 21–24.

[106] D. M. Howard, “Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers”, *Journal of Voice*, vol. 9, no. 2, pp. 163–172, 1995.

[107] I. R. Titze, “Theoretical analysis of maximum flow declination rate versus maximum area declination rate in phonation”, *Journal of speech, language, and hearing research: JSLHR*, vol. 49, no. 2, pp. 439–447, Apr. 2006.

[108] R. Fraile and J. I. Godino-Llorente, “Cepstral peak prominence: A comprehensive analysis”, *Biomedical Signal Processing and Control*, vol. 14, no. 1, pp. 42–54, 2014.

[109] S. Patel, K. R. Scherer, E. Björkner, and J. Sundberg, “Mapping emotions into acoustic space: The role of voice production”, *Biological psychology*, vol. 87, no. 1, pp. 93–98, 2011.

[110] P. Alku and T. Backstrom, “Normalized amplitude quotient for parametrization of the glottal flow”, *J*, vol. 112, no. 2, pp. 701–710, Aug. 2002.

[111] T. Waaramaa, P. Alku, and A.-M. Laukkanen, “The role of f3 in the vocal expression of emotions”, *Logopedics, Phoniatrics, Vocology*, vol. 31, no. 4, pp. 153–156, 2006.

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

[113] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, “Modeling of the glottal flow derivative waveform with application to speaker identification”, *IEEE*, vol. 7, no. 5, pp. 569–576, Sep. 1999.

[114] 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.

[115] M. Huckvale, "Prediction of cognitive load from speech with the VOQAL voice quality toolbox for the interspeech 2014 computational paralinguistics challenge.", 2014.

[116] 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*, Springer, 2014, pp. 275–281.

[117] G. F. Wilson, "Real-time adaptive aiding using psychophysiological operator state assessment", *Publication of: Ashgate Publishing Company*, 2001.

[118] T. Kailath, "The divergence and bhattacharyya distance measures in signal selection", *IEEE Transactions on Communications*, vol. 15, no. 1, pp. 52–60, Feb. 1967.

[119] M. Müller, "Dynamic time warping", *Information retrieval for music and motion*, pp. 69–84, 2007.

[120] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987.

[121] A. Stuiver, K. A. Brookhuis, D. de Waard, and B. Mulder, "Short-term cardiovascular measures for driver support: Increasing sensitivity for detecting changes in mental workload", *International Journal of Psychophysiology*, vol. 92, no. 1, pp. 35–41, 2014.

[122] C. M. MacLeod, "Half a century of research on the stroop effect: An integrative review.", *Psychological bulletin*, vol. 109, no. 2, p. 163, 1991.

School of Science and Engineering
Reykjavík University
Menntavegur 1
101 Reykjavík, Iceland
Tel. +354 599 6200
Fax +354 599 6201
www.ru.is