MY-AHA
Contract # 689592

My-AHA

Deliverable 4.5

Report of validation of My-AHA algorithms

Editor:	Helios De Rosario (IBV)
Deliverable nature:	R
Dissemination level: (Confidentiality)	PU
Contractual delivery date:	M12
Actual delivery date:	M12
Suggested readers:	Developers creating software components to be integrated into knowledge workspace.
Version:	1.0
Total number of pages:	38
Keywords:	Algorithms, sensors, heart rate variability, speech, activity recognition, eye tracking, sit-to-stand power, gait complexity, recurrence quantification analysis, multiscale entropy

Abstract

This deliverable reports the results of the pilot experiments carried out in controlled conditions by IBV, IXP, DSHS and USI to validate six algorithms to process physiological signals, eye-tracking measures and user movements that can be measured with the sensors associated to My-AHA, namely: (1) analysis of heart rate variability, (2) speech analysis, (3) activity recognition by electrooculography, (4) detection of eye movements and blinks, (5) sit-to-stand power, and (6) gait complexity. The results of those tests will be used as a basis for forthcoming decisions about the development of My-AHA platform and its associated modules.

Disclaimer

[Full project title] MY-AHA – **my A**ctive and **He**althy **A**geing

[Short project title] MY-AHA

[Number and title of work-package] WP4, Middleware, Sensing and Data Processing

[Document title] Report of validation of My-AHA algorithms

[Editor: Name, Partner] Helios De Rosario, IBV

[Work-package leader: Name, Partner] Helios De Rosario, IBV

Executive summary

The deliverables D4.2, D4.3 and D4.4 of the project My-AHA present the definition of various algorithms that can be used to process physiological signals, eye-tracking measures and movements of the users, respectively, recorded by sensors associated to the original platforms that will be integrated into My-AHA. Those algorithms were evaluated and prioritised according to their relevance for the interventions that will be delivered by My-AHA, the effort required to implement them in the system, and their scientific and technical interest.

This deliverable reports the results of the pilot experiments carried out in controlled conditions by IBV, IXP, DSHS and USI to validate six of those algorithms, two out of each group of measures, which have been selected according to the abovementioned criteria, namely: (1) analysis of heart rate variability, (2) speech analysis, (3) activity recognition by electrooculography, (4) detection of eye movements and blinks, (5) sit-to-stand power, and (6) gait complexity.

Heart rate variability (HRV) was analysed from the beat-to-beat time intervals captured by the Mio Alpha wristband, in static and dynamic activities (sitting, walking, running), and the results were compared with those delivered by a Polar sensor. The long-term and low frequency features of HRV resulted to be reliable, but short-term and high frequency features could not be successfully retrieved, presumably because of the filters implemented in the acquisition software to reduce the effects of noise and movement-induced artefacts.

The analysis of speech features was used to evaluate the fatigue of people, and predict equivalent scores in the Karolinska Sleepiness Scale. The predicted scores were consistent with the self-reported values using the standard questionnaires, and resulted to be less dependent on the differences accross subjects than relying on their self report.

A machine learning algorithm based on Support Vector Machines was used to analyse the electrooculographic data (EOG) given by the instrumented glasses produced by the Japanese company JINS MEME, in order to discriminate between reading, watching TV and drinking. The algorithm presented a success rate around 86%, much better than popular basic classifiers, depending on the training data and codebooks.

The EOG signals of the JINS MEME glasses were also used to extract eye movements in the four main directions (up, down, left, right) and blinks. The recognition of eye movements was successful, with very high sensitivities and good specificities for the recognition of directional movements. On the other hand, the sensitivity to detect blinks was smaller, since they are faster actions that may be missed due to the limited sampling rate of the sensor.

The muscle power during the sit-to-stand gesture was measured in laboratory by different methods, including high-quality photogrammetry and dynamometric platforms, and a simpler procedure with motion sensors that can be implemented in My-AHA, which yielded values of the average power during the rising phase that were in the same range as the gold standard measures.

Different nonlinear measures of gait complexity, like recurrence quantification parameters and multiscale entropy, were computed for the acceleration signal retrieved from sensors continuously worn by older persons during real-life daily activities. That analysis proved to be feasible in terms of time and computer resources expenditure, in spite of the big amounts of data that have to be processed, and the outcomes resulted to be consistent with previous results reported in literature.

These results will be used as a basis for forthcoming decisions about the development of My-AHA platform and its associated modules.

List of authors

Company	Author
IBV	Helios De Rosario
IBV	Enric Medina
IBV	Elsa Alcarria
IXP	Lennart Kraus
IXP	Jarek Krajewski
DSHS	Eleftheria Giannouli
DSHS	Wiebren Zijlstra
USI	Przemysław Łagodziński

Table of Contents

Abbreviations

CI	Complexity Index
CoM	Centre of Masses
DET	Determinism (RQA parameter)
DIV	Divergence (RQA parameter)
DSHS	Deutsche Sporthochschüle Köln
ECG	Electrocardiogram
ENT	Entropy (RQA parameter)
EOG	Electrooculography
FFT	Fast Fourier Transform
HF	High Frequency
HR	Heart Rate
HRV	Heart Rate Variability
IBV	Instituto de Biomecánica de Valencia
IMR	Interquartile-Median Ratio
IXP	Institute of Experimental Psychophysiology
KSS	Karolinska Sleepiness Scale
L	Average length of recurrent patterns (RQA parameter)
LAM	Laminarity (RQA parameter)
LF	Low Frequency
MFCC	Mel Frequency Cepstral Coefficients
MMSE	Modified Multiscale Entropy
MSE	Multiscale Entropy
NN	"Normal-to-Normal" intervals (intervals between successive heart beats)
RCMSE	Refined Composite Multiscale Entropy
RMSSD	Root Mean Square of Successive Differences
RQA	Recurrence Quantification Analysis
RR	Recurrence Rate (RQA parameter)
SDC	Shifted Delta Cepstral
SDNN	Standard Deviation of NN intervals
STS	Sit-to-stand
SVM	Support Vector Machine
TP	Total Power
TT	Trapping Time (RQA parameter)
USI	University of Siegen
VLF	Very Low Frequency

1 Introduction

One of the key features of My-AHA will be the automatic monitoring and collection of data related to movements, behaviours or other physical or physiological signs, which can be related to frailty symptoms, or may be used to tailor the intervention programs and give feedback about them to the users. Such monitoring will take place by collecting data from different sensors, associated to already existing platforms that will be integrated into My-AHA.

During the first months of the project, the Consortium formed by the research centres, universities and companies involved in the development of My-AHA has also been investigating on methods to enhance the signals and obtain new parameters from the data delivered by the sensors, to feed the analytical models of frailty risk detection and monitor the interventions. That investigation resulted in a collection of algorithms to pre-process physiological signals, eye and body movements, which were presented in the deliverables D4.2, D4.3 and D4.4.

The algorithms and associated measures presented in those deliverables have been evaluated during content and technical meetings of the project Consortium, taking into account their relevance for monitoring different types of intervention, and the complexity of implementing them in My-AHA and/or the original platforms. Tables 1, 2 and 3 present the full list of algorithms that were described in those deliverables, with the related frailty domains and the main highlights of the evaluation carried out by the Consortium.

Table 1. Algorithms proposed to process physiological signals (D4.2)

Algorithms	Domains	Comments
Cardiac feature extraction	Physical, cognitive, sleep, psychological	Heart rate implemented in Beddit & wristbands like Mio Alpha. No variability calculations implemented in the devices. It can be used in the physical interventions (cardiovascular training).
Respiration feature extraction	Physical, sleep, psychological	Breath lengths and frequencies implemented in Beddit.
Speech feature extraction	Cognitive, psychological	Currently implemented in desktop application; to be implemented as mobile app for My-AHA.

Table 2. Algorithms proposed to process eye movements (D4.3)

Algorithms	Domains	Comments
Analysis of eye movements and blinks	Cognitive	Attainable from JINS MEME glasses. Validation and comparison with gold standard is needed.
Measurement of fatigue and drowsiness	Physical, sleep	Measured by blink rate and speed from JINS MEME glasses. It depends on the development of eye movement/blink detection.
Machine-learning classifiers	Physical, cognitive	Measured by the EOG signal from the JINS MEME glasses.

Table 3. Algorithms proposed to process user movements (D4.4)

Algorithms	Domains	Comments
Step segmentation	Physical	Measured from accelerometer signals, currently not implemented. To be validated against gold standard.
Measurement of walking activity	Physical, sleep, psychological	High priority for interventions. Fully implemented in Smart Companion and Medisana activity trackers.
Gait speed	Physical	High priority for interventions. Fully implemented in Smart Companion and Medisana activity trackers
Gait variability	Physical	High priority for interventions. Fully implemented in Smart Companion
Gait complexity	Physical	Measured from accelerometer signals, currently not implemented. Must verify that the computational resources needed to implement it are feasible.
Medial-lateral control	Physical	Measured from accelerometer signals, not implemented. To be validated against gold standard.
Unassisted tests (Sit-to-stand, TUG, One Leg, sway)	Physical	Partially implemented in modules that can be associated to the Smart Companion / Smart Feet. Highest priority among missing features given to the measure of Sit-to-stand power.
Trend Analysis	Physical	Developed for iStoppFalls movement data.

For each group of variables (physiological signals, eye tracking, and user movements) we have selected two algorithms to be validated and further developed. This selection has been done on the basis of the following criteria:

1. Relevance for the interventions delivered by My-AHA.
2. Feasibility of retrieving the required signals from one or more devices that can be integrated into My-AHA.
3. Complexity of implementation.
4. Scientific-technical interest of implementing and testing it for real-life monitoring (still not present in the commercial devices).

The algorithms that have been selected are:

- For the analysis of physiological signals:
    - Heart rate variability measures from the Mio Alpha wristband sensor.
    - Speech feature extraction during voice interaction to evaluate the level of fatigue/tiredness.
- For eye tracking (with the JINS MEME glasses):
    - Activity classification from EOG data.
    - Detection of gaze movements and blinks.
- For user movements:
    - Sit-to-stand power (muscle strength).
    - Gait complexity.

This deliverable reports the results of the pilot experiments carried out by IBV, IXP, DSHS and USI to validate those methods. Those results will be used as a basis in forthcoming decisions for the development of My-AHA platform and its associated modules.

2 Heart rate and heart rate variability

2.1 Objectives of the validation

Wrist-worn trackers like Mio Alpha are focused on the correct detection of heart pulses to measure heart rate (HR). That parameter is useful to monitor cardiovascular training, which is often the purpose of such devices. But in relation with the ageing process, the analysis of heart rate variability (HRV) is also relevant, since it can quantify the loss of autonomic influences on HR regulation as a function of age (Antelmi et al. 2004; Lipsitz et al. 1990).

There are many estimators of HRV, based on time or frequency domain features of the NN interval time series (the sequence of beat-to-beat time intervals). Among time domain features the most common are:

- The standard deviation of NN intervals (SDNN), related to the overall variation of HR along the whole record.
- The root mean square of successive differences (RMSSD), related to the immediate variation of NN intervals.
- The percent of successive NN intervals that differ more than 50 ms from each other (pNN50).

Frequency domain features are based on the power of the NN time series within different frequency bands:

- Total power of the signal (TP).
- Power in the "very low frequency" band (VLF), below 0.04 Hz.
- Power in the "low frequency" band (LF), between 0.04 Hz and 0.15 Hz.
- Power in the "high frequency" band (HF), between 0.15 Hz and 0.4 Hz.

There are also normalised measures of the frequency domain features, based on ratios between the previous parameters.

The frequency spectrum of the NN intervals is normally computed through the Fast Fourier Transform (FFT) of the time series, but that operation assumes that the signal is sampled at fixed intervals, which is not the case of the observed sequence of NN intervals, so for a strict computation of frequency domain features it is necessary to interpolate the observed sequence in an evenly spaced time line, at a frequency of 0.4 Hz or higher. However, for faster computations the FFT can be applied directly to the observed time series, assuming some level of error in the results.

In the case of NN intervals measured by wrist-worn sensors like Mio Alpha, we can assume that there are other sources of error, derived from motion artefacts or the data filtering and other pre-processing that must be applied to the sensor data to rule out those artefacts. Therefore, it is questionable whether the HR and HRV parameters obtained from them are reliable, and if the complexity added by the interpolation makes any difference. The validation that has been carried out attempts to give an answer to those two particular questions.

2.2 Material and methods

HR and HRV were measured with the wrist-worn Mio Alpha sensor, and compared with a chest-worn Polar sensor during controlled experiments carried out in IBV. 10 young adults (5 male and 5 female, aged between 25 and 45) participated in experimental sessions, in which their cardiac activity was measured with those two sensors simultaneously, during two tasks:

1. Sitting quiet during approximately 4 minutes, with controlled arm movements.
2. Treadmill walking/running in 3 bouts at a progressively increasing speed: 4, 6, and 8 km/h. Each bout was 4-minutes long, with resting periods of 3 minutes at the beginning, at the end, and between each two bouts.

The NN intervals were extracted in real time from the Mio Alpha device by streaming the data via Bluetooth with the Cardiomood app (www.cardiomood.com), and from the Polar logs recorded by the Polar Protrainer application (www.polar.com). Polar data was considered as gold standard, as it has been demonstrated to be highly reliable compared with standard ECG (Nunan et al. 2008; Vanderlei et al. 2008; Weippert et al. 2010).

The series of NN intervals taken from both devices were analysed both in the time and frequency domains, and the outcomes were compared to evaluate the similarity between HR and HRV measures obtained from each device. The analysis in the frequency domain was done with and without interpolation (using cubic splines), to compare the loss of accuracy that might be expected if the interpolation step is ruled out.

Since the distributions of HR and HRV are typically right-skewed, the similarity between the results of both devices was evaluated as differences in a logarithmic scale, i.e. of ratios between HR and HRV parameters, instead of differences.

2.3 Results

2.3.1 Collected data and missing values

The Polar sensor was firmly pressed against the user's chest, and allowed a perfect detection of heart beats in the quiet measures, except in one user for which a significant amount of data was lost. The Bluetooth connection of Mio Alpha was temporarily interrupted in 80% of the measures, although in most cases the interruptions were short enough to be able to carry out the analysis, ruling out the data points before and after each interruption, which was detected by outliers in the distribution of the NN time series. All in all, it was possible to compare 15 out of the 20 records that were taken: 9 of the sitting task (ruling out the case with the loss of Polar data), and 6 of the walking/running task (ruling out 4 measures with significant data loss of Mio Alpha, which precluded a reliable comparison with the Polar records).

2.3.2 Comparison of HR and HRV parameters

The NN intervals recorded by Mio Alpha were highly correlated with those measured with Polar, although they were heavily smoothed, presumably due to the filters applied to cancel motion artefacts (see Figure 1). That smoothing had an important influence on HRV measures, specially those that assessed the short-term variability (in the time domain) and high frequency components (in the frequency domain).

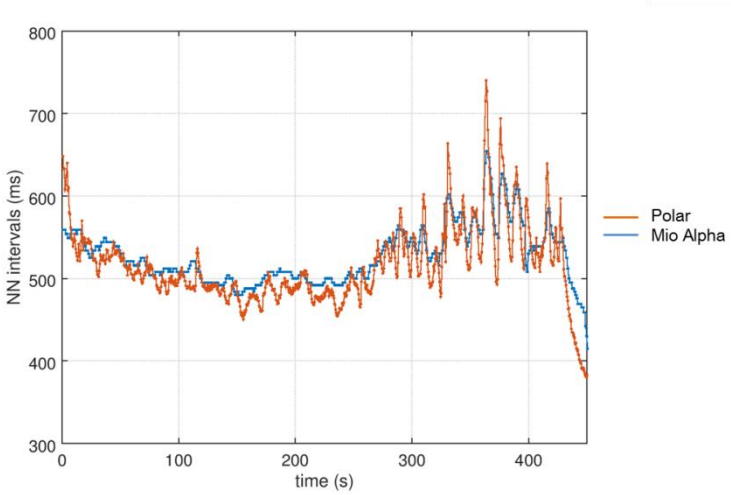

Figure 1. Example fragment of NN time series

That influence is clearly seen in the plots of Figure 2 and Figure 3. Those figures show side by side the paired values of HR and HRV resulting from analysing the NN sequences with each device, and the ratios between those values (Mio w.r.t. Polar), with confidence intervals around the linear trends of the ratios as a function of the values measured with Polar.

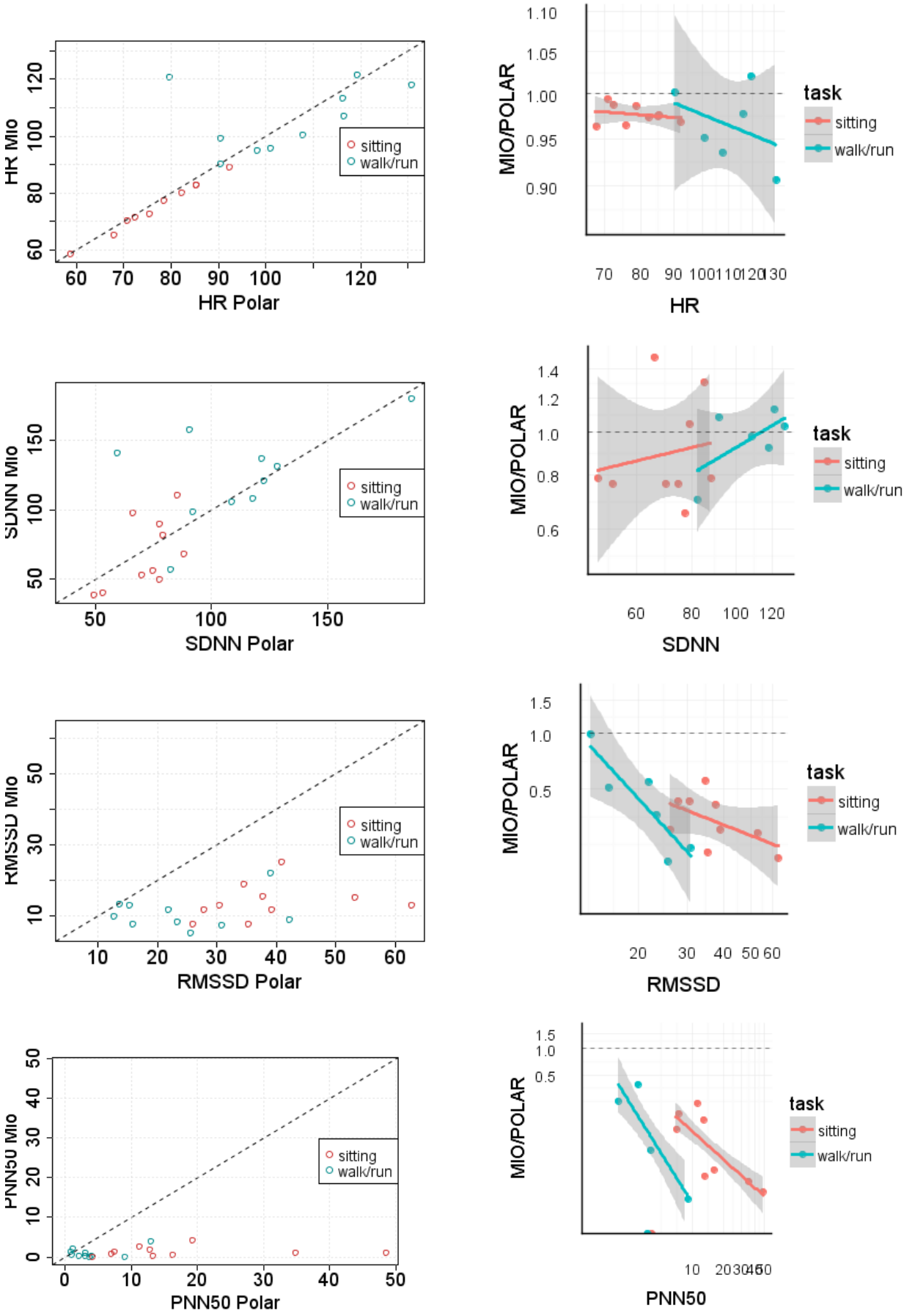

Figure 2. Comparison of HRV time-domain features

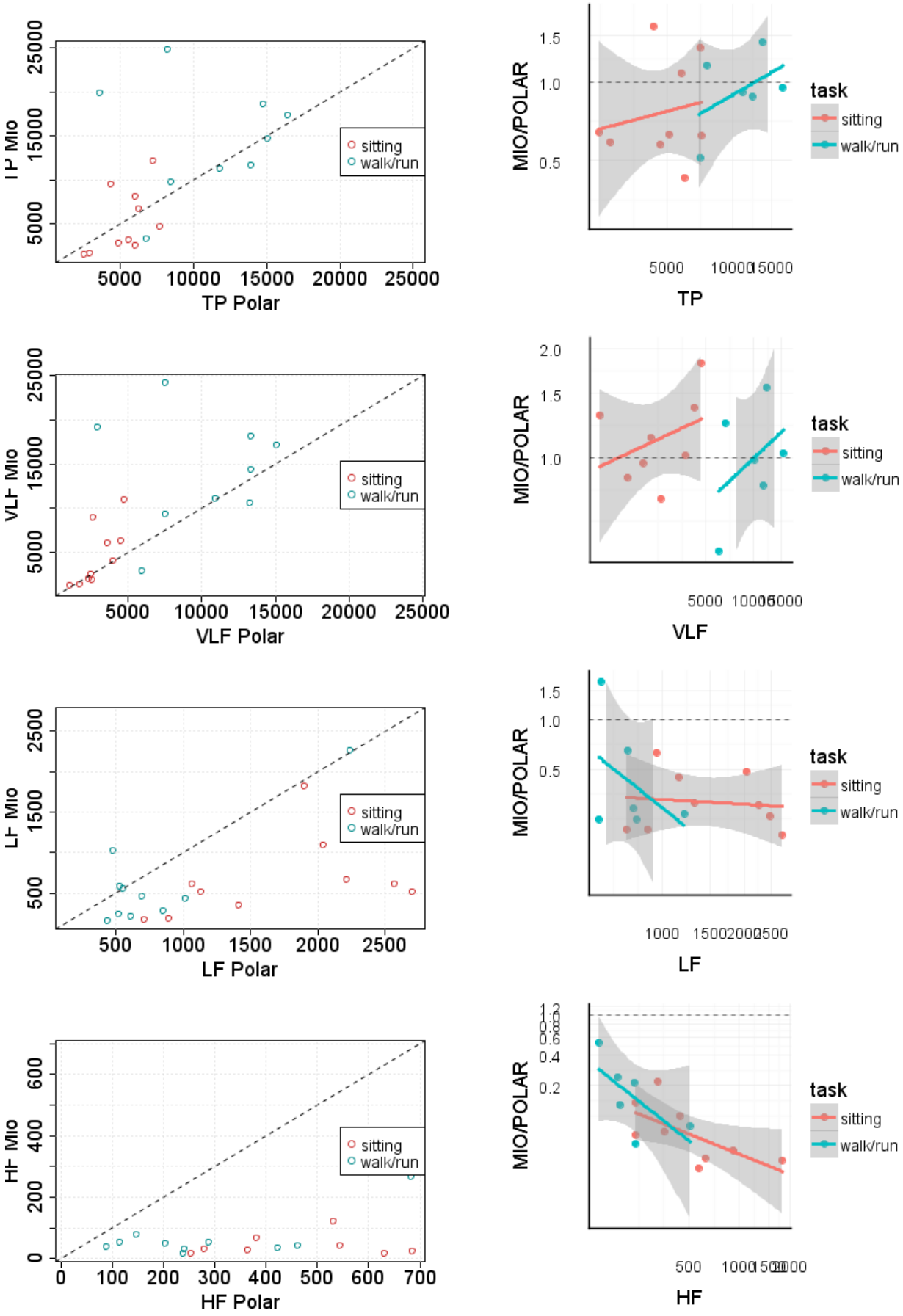

Figure 3. Comparison of HRV frequency-domain features

It may be seen that HR values computed from Mio Alpha data were comparable to the gold standard measured by Polar. There was a small bias towards smaller values, but the overall ratio of Mio vs. Polar data was very close to the unit (between 0.91 and 1.02, with greater dispersion when the users were moving).

The ratios of SDNN values were also around the unit, albeit with greater dispersion. On the other hand, RMSSD and pNN50 values were clearly underestimated. The decreasing linear trend of the ratios between Mio and Polar for those features are due to the fact that the values computed from Mio Alpha data were around a fixed range of values (approximately between 5 and 25 ms for RMSSD, and less than 10% for pNN50), regardless of the actual variations of that feature. Thus, although those HRV parameters should be smaller during walking and running compared to sitting (Chan et al. 2007), as observed with Polar, those variables did not show differences between tasks when they were calculated from Mio Alpha data.

A similar situation happened with the frequency domain features of HRV: the power calculated from the data of Mio alpha in the whole spectrum of the signal (TP) and in the VLF band varied following a trend that was coherent with the measures taken with Polar, although there was a high dispersion, as happened in SDNN. In fact, TP and VLF provided basically the same information as SDNN for Mio Alpha data, since nearly the whole signal was contained in the VLF band (over 75% of power in all measures). On the other hand, the power of the LF and specially the HF band was severely underestimated.

2.3.3 Influence of interpolation in frequency domain features

As shown in Table 4, the frequency domain features of HRV calculated without interpolating the NN sequences were similar to the outcomes obtained from the interpolated time series (ratios between the results without interpolation vs. interpolated near to the unit). However, there was a significant bias towards higher ratios, i.e. the calculated powers were higher when the interpolation step was neglected, except in the LF band.

Nevertheless, that bias is insignificant compared with the high errors in the LF and HF components that are reported in the previous section. On the other hand, the information contained in the TP and VLF bands is virtually equivalent to the results obtained from SDNN in the time domain, so all in all, the subtle differences between procedures to calculate frequency domain features can be considered to be irrelevant for practical purposes.

Table 4. Statistics of the power of the NN signal in different frequency bands, including ratios of the power calculated without signal interpolation vs. calculations with the interpolated signal

	Mean power	std. dev.	mean ratio	Conf. interval of the ratio*	t*(14)	p-value*
TP	7160.207	5290.330	1.073	1.007 - 1.132	2.408050	0.030
VLF	6634.192	5286.604	1.078	1.005 - 1.142	2.309880	0.036
LF	475.266	280.506	1.042	0.980 - 1.360	1.359964	0.195
HF	39.051	20.037	1.101	1.030 - 1.181	3.180736	0.007

(*) The statistical analysis is done for the differences of the logarithms of TP, VLF, LF and HF values. The confidence intervals of the ratios are calculated by exponentiation of the confidence intervals for the differences between logarithms.

3 Speech analysis

3.1 Objectives of the validation

The focus of the speech analysis tool is to extract features from the speaker's audio signal which may indicate emotional or physiological states. This kind of parameters is a useful tool to monitor the state of the user without attaching anything directly to the body. This may also be implemented without obtrusive microphones as long as adequate noise reduction is applied.

A prominent component of frailty is fatigue, which can be quantified through the Karolinska Sleepiness Scale (KSS) score. The KSS is a 9-point self-reported, verbally anchored scale that measures the subject's fatigue, going from 'extremely alert' to 'extremely sleepy/ fighting sleep', which is closely related to encephalographic and behavioural indicators of alertness/sleepiness (Kaida et al. 2006), and it is frequently used in studies measuring subjective fatigue.

The experiment presented in this report had the purpose of testing the validity of speech analysis as an objective tool to assess the fatigue of subjects interacting verbally with a computer-based application, compared to subjective assessments as measured by the KSS.

3.2 Material and methods

3.2.1 Set up and instruments

For the identification of vigilance-induced phonetic-linguistic changes in the speech and the associated ascertainment of the necessary vocal data for the estimation of the available vigilance scoring, a corresponding technical setting was implemented in the control room simulation environment at Psyrecon GmbH by RFH Cologne. For this purpose a headset as well as a computer with corresponding recording software was provided, and a screen presentation was created, which could be controlled by the Proband experimental director (Figure 4). The purpose of this test arrangement was to record the speech samples of the subjects before and during the main task, to provide them with time stamps for each subject, and to store them together with a fatigue score estimated by the subject.

Questionnaires were also drawn up for recording third-party variables. Variables that can influence vigilance, such as gender, age, height, weight, smoking status, cold status, regional dialect, and verbal intelligence were recorded. In the course of the acquisition, attention was paid to the age of the subjects and to an adapted diversity with regard to age class and gender.

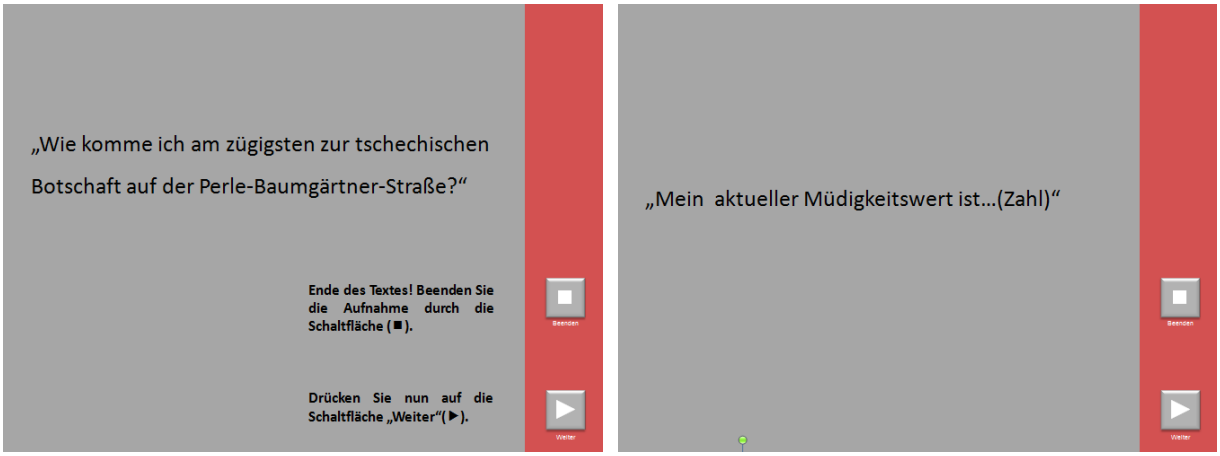

Figure 4. Graphical interface used in the speech analysis experiment

3.2.2 Subjects and tasks

Measures were taken for 38 German-speaking adults, evenly distributed between male and female. Each subject participated in various sessions at different times of the day having slept normally and with deprivation of sleep. We recorded between 37 and 44 measures of the KSS scores for each subject, both verbally reported and predicted by the speech analysis.

The speech tasks were composed of reading tasks (phonetically balanced texts such as "The North Wind and the Sun" as well as "The Butter Story") and sentences like "How do I get to the Czech Embassy on Perle-Baumgärtnerstrasse?" in German language. The verbal assessment of the fatigue of the subject as a self-report on the KSS ("My current tiredness feeling is at x") was also recorded. Additional language material was gained by tasks from the Thematic Conviction Test (free picture descriptions).

During the recording of the speech samples, the experimenter and the subject were connected via a headset. The sentence "How do I get to the Czech Embassy on the Perle-Baumgärtnerstrasse?" was recorded in 4 minute intervals during the main task. The fatigue questionnaire "My current fatigue feeling at x" was recorded in 8 minute intervals. The fatigue score reported by the subject was entered directly after the speech task by the researcher.

3.2.3 Speech analysis

There are many state estimation parameters that may extracted from a speech vector. Common speech analysis features are:

- Mel Frequency Cepstral Coefficients (MFCC). MFCC are coefficients that collectively form Mel frequency cepstrum which is a power spectrum of a short window of a speech signal (i.e. 40 milliseconds). MFCC tries to represent the shape of the vocal tract using the short term power spectrum thus trying to approximate human auditory system responses.
- Shifted Delta Cepstral (SDC). Shifted Delta Cepstral coefficients are considered to be long term features which are derived using MFCCs. SDC has 4 parameters (*N*, *d*, *p*, *k*), described as follows:
    - *N* is the number of cepstral coefficients computed at each frame
    - *d* represents the time advance and delay for the delta computation
    - *P* is the time shift between consecutive blocks
    - *k* is the number of blocks whose delta coefficients are concatenated to form the final feature vector
- Pitch. The pitch or fundamental frequency is considered to be a most basic discriminating factor between the male and female voice.

In order to ensure standardization, the audio data used to acquire features mentioned above needs to be uniform in structure. For this reason the structure given for the audio signal to produce all extracted features was set to:

- One second in length
- Mono channel
- 44.1 kHz sampling frequency
- 16 bit data resolution

The extracted features were used in conjunction with a machine learning model to estimate the fatigue score which correlates to the audio sample provided. The model used for the estimation had first to be trained on a database of audio files with associated KSS scores. This allowed for a large training set to be used for the estimation process, and was more likely to provide a more accurate estimation result. The model used can be from various structures such as:

- Linear Regression Model
  http://eli.thegreenplace.net/2014/derivation-of-the-normal-equation-for-linear-regression/
    - The hypothesis function for a linear model is given as:

- $h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n$
- Assuming that the matrix $X^T X$ is invertible, we can simplify the calculation of the vector θ such that: $\theta = (X^T X)^{-1} X^T y$

- Support Vector Machine (SVM)
  http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html
  - Support Vector Machines are discriminative classifiers formally defined by a separating hyperplane. In other words, given labelled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.
  - The notation used to define formally a hyperplane is:
  - $f(x) = \beta_0 + \beta^T x$, where β is know as the weight vector and β_0 as the bias.
- Gaussian Mixture Model
  http://www.ee.iisc.ac.in/people/faculty/prasantg/downloads/GMM_Tutorial_Reynolds.pdf
  - A Gaussian Mixture Model is a parametric probability density function represented as a weighted sum of Gaussian component densities. They are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system.
  - A Gaussian mixture model is a weighted sum of M component Gaussian densities as given by the equation:
  - $p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i)$, where x is a D-dimensional continuous-valued data vector (i.e. measurement or features), w_i, $i = 1, \ldots, M$, are the mixture weights, and $g(x|\mu_i, \Sigma_i)$ are the component Gaussian densities.

3.2.4 Statistical analysis

The KSS scores obtained by subject's answers to the questionnaires and the scores predicted by the speech analysis were analysed, considering that they might depend on the state of the subjects (awake or tired), and that there might be a considerable variability across individuals. To account for those sources of variability, the following linear mixed model was fitted to the resulting data:

$$y_{sij} = \mu + \alpha_s + \eta_i + \epsilon_{ij},$$

where:

- y_{sij} was the outcome of interest (the actual KSS score or its difference with the predicted one) for the measure j of the subject i in state s,
- μ is the average expected value in the control state (awake),
- α_s is the expected increment on that average in the tired state, and
- η_i, ϵ_{ij} are random effects due to the inter-subject and within-subject inherent variability to the measure, respectively.

The variance of those random effects, and the expected scores and differences were analysed by an ANOVA of that model.

3.3 Results

3.3.1 Comparison of estimated KSS and supplied KSS

Figure 5 shows the speech signal and its Mel frequency cepstrum of a person telling the word "tschechische" ("Czech" in German), both in the awake and tired states.

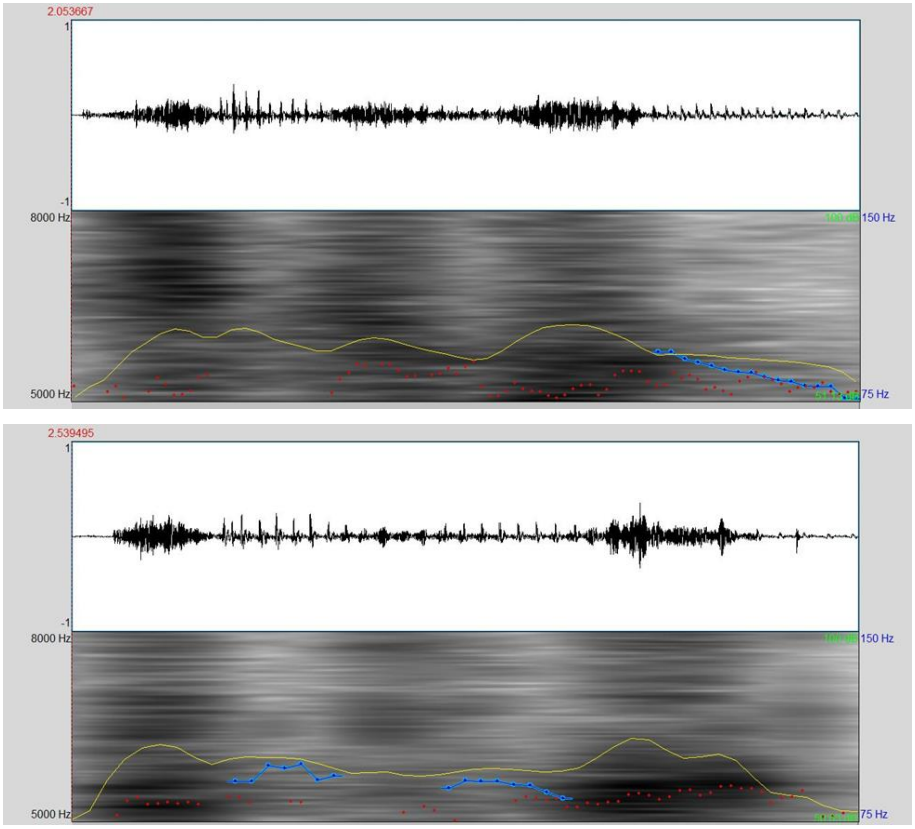

Figure 5. Example of awake person (up) and tired person (down) telling "tschechische"

The reported KSS scores varied between the whole range of the scale (1 to 9). In the awake state the average score was around 4 ("rather alert"), with a standard deviation of 1.6 points; in the tired state the subjects tended to report higher levels of fatigue, such that the average value was slightly increased to 4.2, with the same standard deviation.

The ANOVA on the actual KSS scores (Table 5) shows that such increase in the KSS score, albeit small, was statistically significant ($p < 0.02$), so the level of fatigue induced to the participants could be considered to be noticeable. The variability between subjects was approximately the same as the variability observed in the answers of the subjects in the same state ($\sigma_\eta = 1.16$, $\sigma_\epsilon = 1.11$).

Table 5. ANOVA of the actual KSS reported by the subjects

	Value	(std. err)	F(1,1582)	p-value
μ (expected value for awake)	4.06	(0.19)	466.24	0.000
α_s (tired - awake)	0.13	(0.06)	5.45	0.020

The KSS scores predicted by the speech analysis were around the same average values as the reported scores in both states, although the standard deviation was slightly reduced to 1.4. That decrement of variability was mainly due to more consistent predictions between subjects ($\sigma_\eta = 0.85$), whereas the variability within subjects remained the same as for the reported KSS scores.

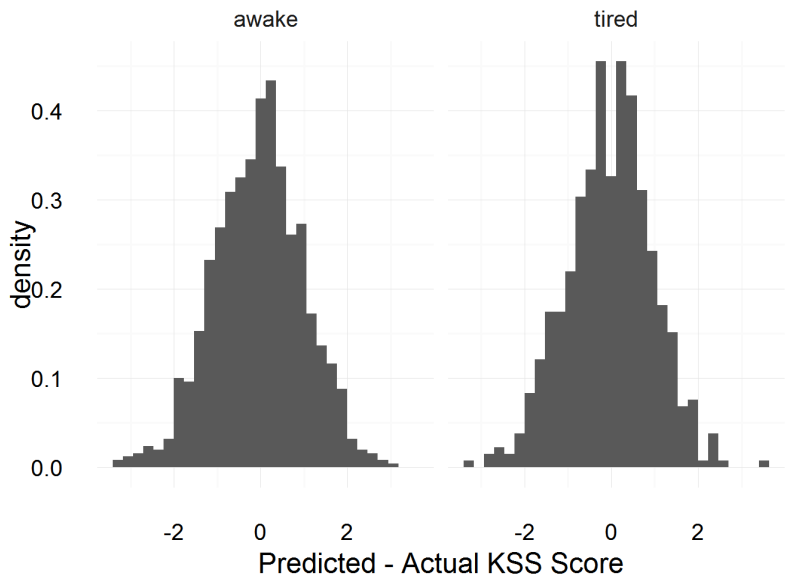

Figure 6. Differences between predicted and actual (reported) KSS scores

Figure 6 shows the distributions of the differencse between actual and predicted scores. It may be observed that the differences were normally distributed in the same fashion for both states (awake and tired).

The ANOVA (Table 6) shows that there is no significant difference between the scores in any of the states (the *p* value is exceedingly greater than 0.05 for the expected difference in the awake state, and also for the expected increment in the tired state). Thus, all differences between predicted and actual KSS scores could be regarded as random, unbiased errors. The standard deviation of such error was $\sigma_\epsilon = 0.94$ within subjects (about the same order of magnitude as the within-subjects variability of the actual KSS score), with an added error between subjects with $\sigma_\eta = 0.38$, which was smaller but still significant ($p < 0.01$).

Table 6. ANOVA of the difference between predicted and actual KSS

	Value	(std. err)	F(1,1582)	p-value
μ (expected difference for awake)	-0.040 (0.07)	(0.07)	0.338	0.561
α_s (tired - awake)	0.005 (0.05)	(0.05)	0.012	0.912

4 Activity recognition with EOG data

4.1 Objectives of the validation

Recently, mobile devices equipped with different sensors have become very popular and accompany us in everyday activities collecting various kind of data. In case of eye-trackers, the availability of wearable devices that do not cause discomfort in human daily activities is somehow limited.

The Japanese company JINS MEME came out with a project of intelligent glasses that comprise electrooculography (EOG) sensors along with built-in accelerometer and gyroscope. The EOG sensors, three electrodes placed around the nasal bridge provide the electrooculogram depicting the eyes movement with a series of values measured as a difference between the electric potential fields of the poles at right and left nose pads (respectively for the right and left eye) and the reference one of the pole located at the bridging part. Based on those two signals (EOG_L and EOG_R) another two are formulated to describe horizontal and vertical eye movement (EOG_H and EOG_V).

Since the obtained information is a low-level sensor data expressed as a sequence representing values in constant intervals (depending on the JINS MEME application settings it is 50, 100 or 200Hz), the human activity recognition problem can be formulated as sequence classification. In order to categorize sequences of sensor values into appropriate activity class various machine learning classifiers were investigated:

- Mean values,
- Standard deviation,
- Mean of absolute values of first differences,
- Mean of absolute values of second differences,
- Feature learning approach called “Codebook approach”.

4.2 Materials and methods

The EOG consisting of four data vectors (EOG_L, EOG_R, EOG_H, EOG_V) along with the accelerometer data were collected with the JINS MEME glasses during experiments in controlled environment at the laboratories of the University of Siegen. 100 adults, mostly the University students, participated in experimental sessions, in which their eye movements were recorded while performing one of the following three daily activities:

a) reading a page of text in a participant native language,

b) drinking mineral water,

c) watching television.

The data set obtained for each of these activities was around 30 seconds long. Since the sampling rate varying from 100 to 200 Hz is considered as high, the amount of data collected within that timeframe should be sufficient enough to be able to build an effective classifier. After each activity, participants took a short brake to calm their eyes.

The EOG data and the accompanying accelerometer values were collected in real time from the JINS MEME glasses by the streaming the data via Bluetooth dongle with the application “JINS MEME Data Logger” provided by the manufacturer of glasses.

The obtained data was divided into several train and test data sets and used in during experiments to evaluate the effectiveness of models based on different classifiers in the manner of activity recognition using SVM for machine learning.

During the experiments another problem became visible in the obtained data. In case of reach facial mimics or when someone was not used to wearing glasses and tried to correct their position a lot of noise was introduced to the EOG signals. This is probably the outcome of the electrodes locations, which are very close

to each other – a less noisy signal could be obtained when reference electrode was located at the nasal bridge and the other two at the outer sides of each eye.

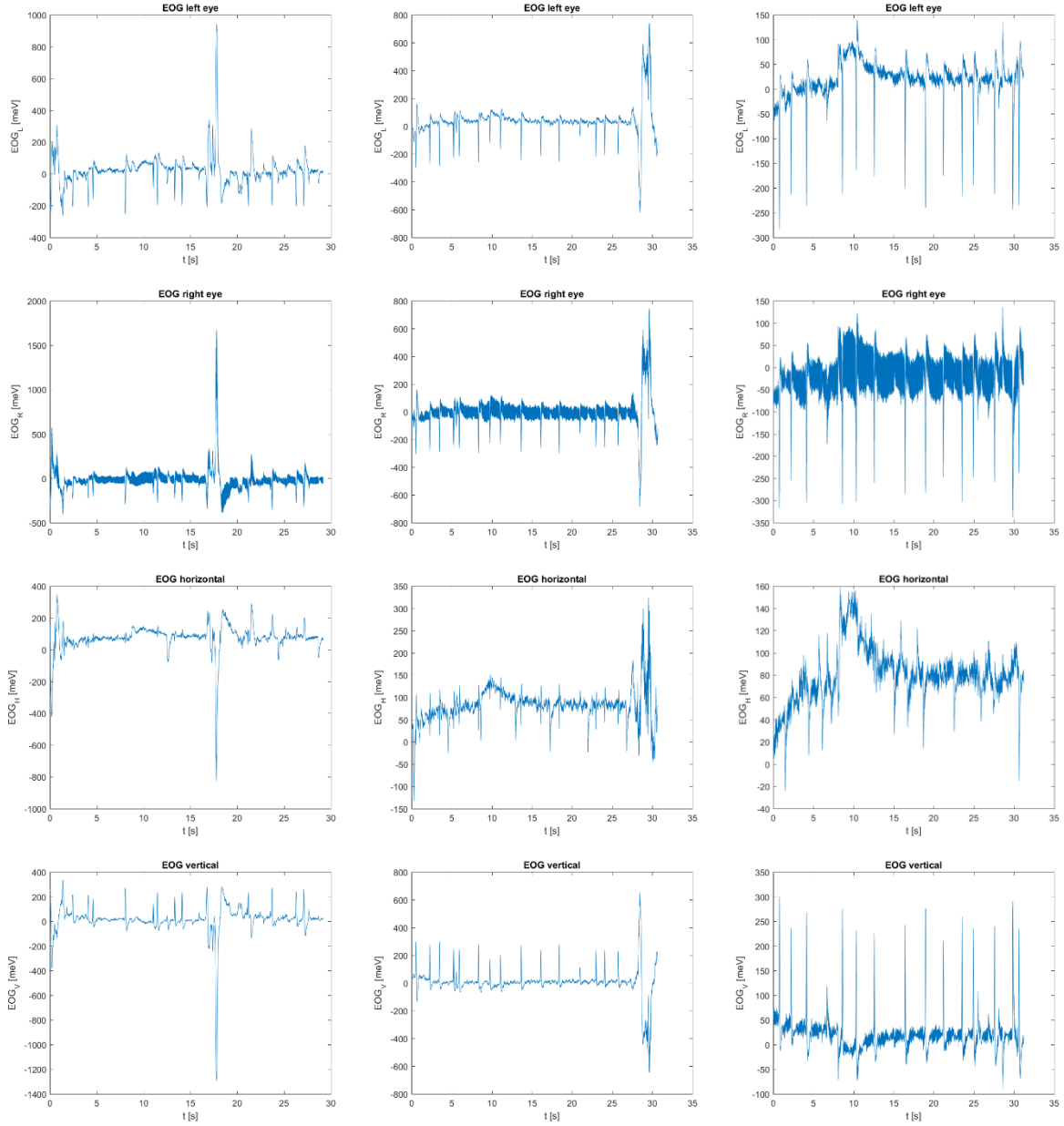

Figure 7. Sample plots depicting the EOG signals collected during experiments. Activities in columns from the left: drinking, reading and watching television

4.3 Results

The first experiments involved the popular basic classifiers: mean values, standard deviation and the mean of absolute values of first and second differences. Results obtained with these classifiers were average. The best results were produced by the mean of absolute values of first and second difference (72,6% and 70,6% respectively), however the cross-reference with changed data sets produced the results around 50%.

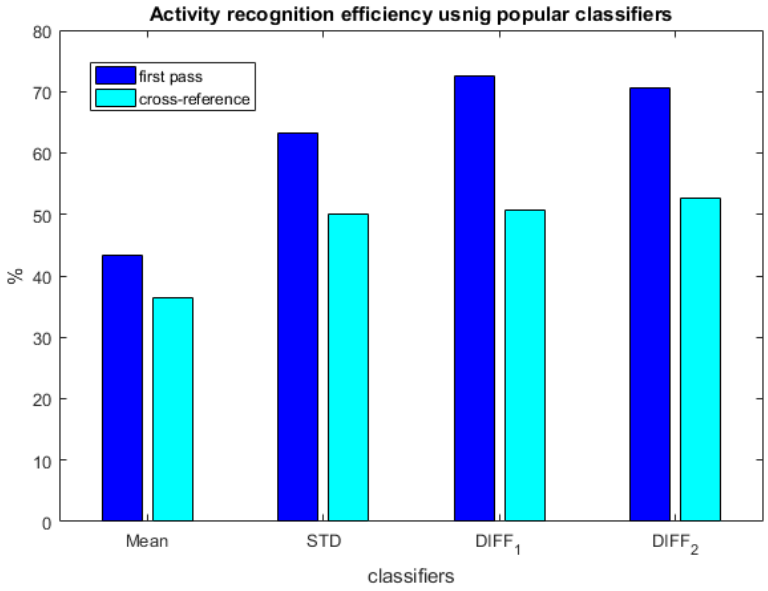

Figure 8. Efficiency of activity recognition using popular classifiers: mean values (Mean), standard deviation (STD), mean absolute values of first differences ($DIFF_1$) and mean absolute values of second differences ($DIFF_2$).

Other experiments were focused on the application of the Codebook approach in order to extract useful features from the collected EOG data. This solution divides sequences into subsequences that are grouped into clusters, called "codewords", characterizing a statistically distinctive subsequence. Next step provides the sequence encoded as a feature representing the distribution of codewords. The Codebook approach produced improved and more consistent results reaching the level of 86% accurately classified activities.

A slightly modified version of the codebook approach using the Fourier coefficients describing consecutive subsequences to build codewords was also investigated. This approach introduced a little of improvement conquering the result to 87,3%.

Further work will focus on combining the feature vectors produced by codebook approach with the feature vectors obtained with classifiers presenting more global characteristics of the data sequence. Also, in case of codebook approach it is also worth investigating if the histograms could be replaced with other distribution models like Gaussian Mixture Model.

In the matter of additional noise introduced to the EOG signals by external factors several experiments were performed in order to investigate if pre-processing actions like smoothing, denoising and approximating the obtained EOG signals could improve the results. In most cases the pre-processing of the data does not improve the activity recognition accuracy, and in case of approximation the results were even worse than with original, noisy data.

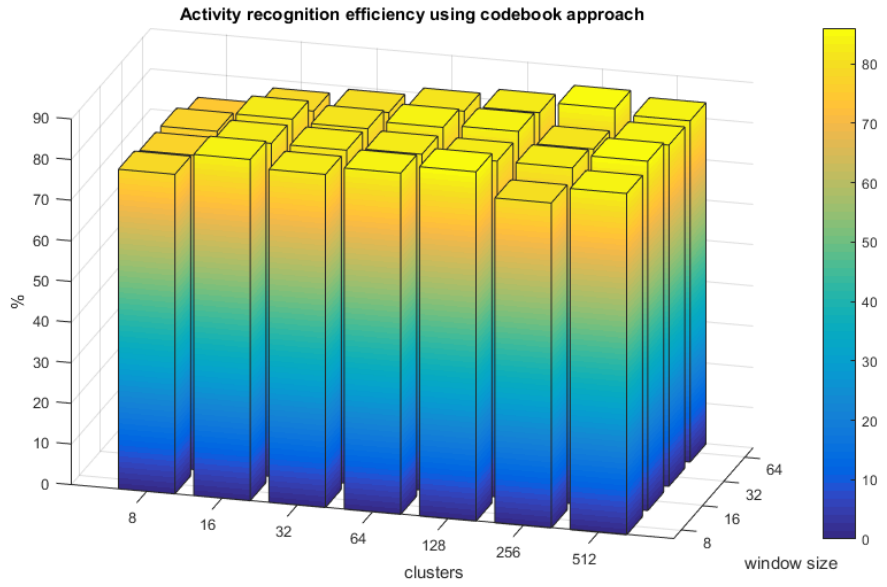

Figure 9. Efficiency of activity recognition using the codebook approach to build the feature vectors. Best result 86% obtained for subsequences of size 8 and 128 clusters.

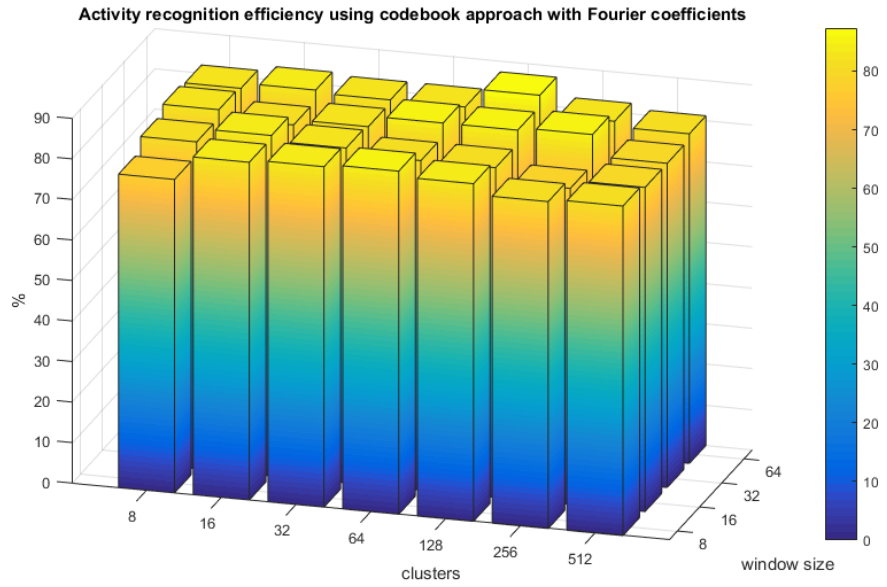

Figure 10. Efficiency of activity recognition using the codebook approach utilizing the Fourier coefficients to build codewords and finally the feature vectors. Best result 87,3% obtained for subsequences of size 64 and 128 clusters

5 Eye movement extraction

5.1 Objectives of the validation

The EOG signals of the JINS MEME glasses can also be used to identify eye saccades in the four main directions (up, down, left, right) and blinks during quiet activities like reading, watching images, etc. Such features may be relevant to assess the cognitive activity of the elders, since changes in the duration of eye fixations, refixations, and saccade inhibition (lack of eye movements) are symptomatic of cognitive decline (Pereira et al. 2014). The experiment that is reported in this chapter was aimed at validating the reliability of a wavelet-based method to classify such movements that was presented in D4.4.

5.2 Materials and methods

Five young adults participated in a controlled experiment to verify the efficacy of the wavelet-based algorithm to detect gaze movements in horizontal (left-right) and vertical (up-down) directions and blinks. The subjects wore the academic version of the instrumented glasses, and sat in front of a computer screen in the laboratories of IBV. The JINS MEME data logger was started at the same time as the Tobii T120 eye tracker, which was used as auxiliary measure to verify the movements of the eyes and blinks.

The subjects were instructed to look at varying directions when they heard an acoustic signal, that was triggered approximately every 3 seconds, during a period of 1 minute. A total between 18 and 28 movements were recorded for each user, distributed across the four directions so that there were between 2 and 7 movements in each direction. The subjects were allowed to blink freely, so that the number of blinks was not fixed.

The number of true and false events (eye movements or blinks) detected by the algorithm was recorded, as well as the events missed or wrongly classified. The sensitivity and specificity of the algorithm was calculated by the following parameters:

- Sensitivity: ability to identify events. It is quantified as the ratio between correctly detected events and actual number of events of that type:

$$Sensitivity = \frac{\#True\ positives}{\#Actual\ events}$$

- Specificity: ability to rule out events of another type or false events. It is inversely related to the ratio between such false or incorrect events and the total number of events detected of one type:

$$Specificity = 1 - \frac{\#False\ positives}{\#Detected\ events}$$

5.3 Results:

In addition to the events triggered by the acoustic signal, each subject blinked between 1 and 4 times during the minute of the test. The vertical and horizontal channels of the EOG signal were extracted and analysed by an 8-level wavelet decomposition (see an example in Figure 11 and Figure 12), and the intermediate and 8th levels were analysed as described in D4.4.

The confusion table presented in Table 7 shows how many events were correctly and wrongly identified for each direction of movement and for blinks. The sensitivity and specificity of the algorithm is reported for each type of event in Table 8.

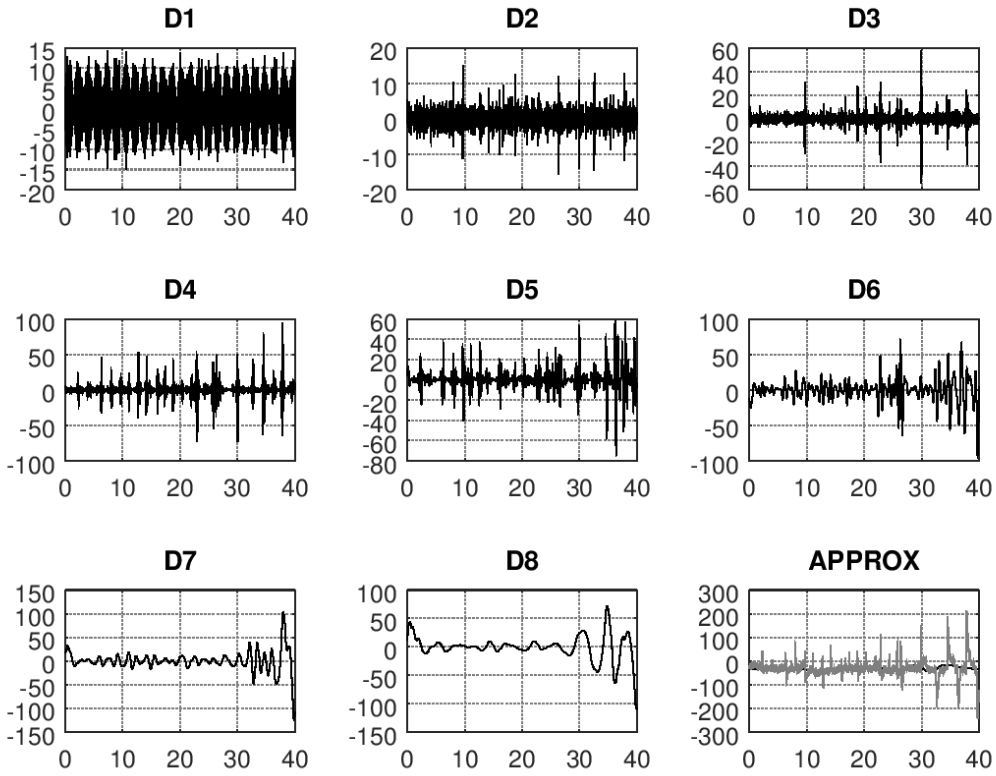

Figure 11. Wavelet decomposition of the vertical EOG signal

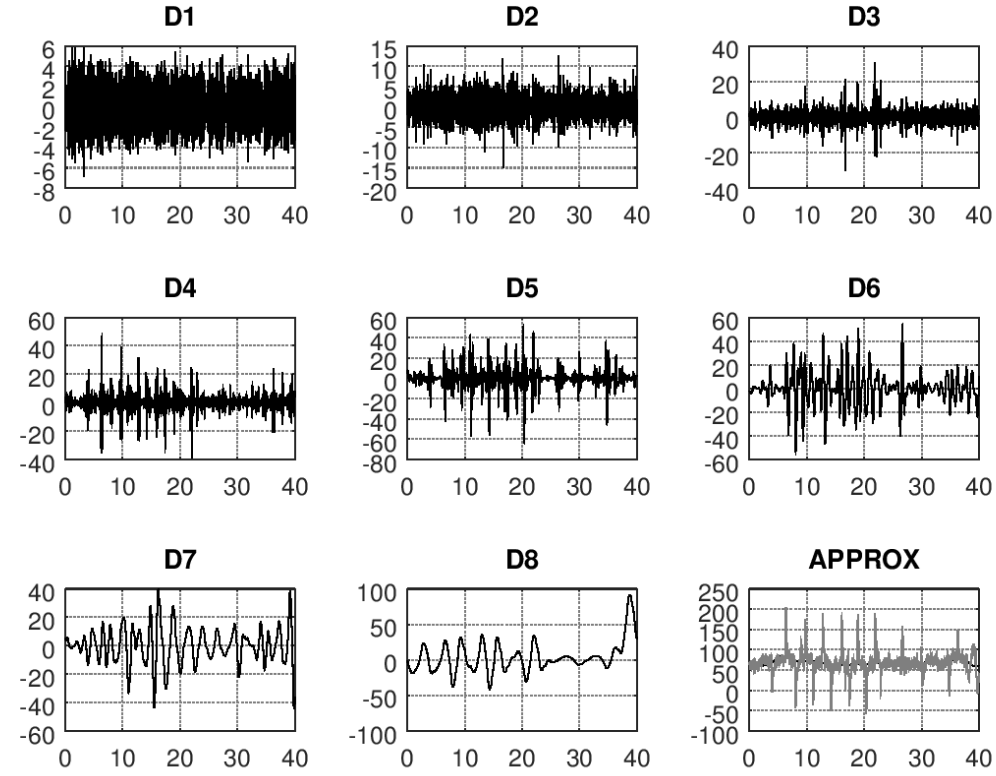

Figure 12. Wavelet decomposition of the horizontal EOG signal

Table 7. Confusion table

		Detected events					
		Up	**Down**	**Left**	**Right**	**Blink**	**No detection**
Actual events	**Up**	23	0	1	1	0	0
	Down	0	29	2	0	0	0
	Left	0	0	23	0	0	0
	Right	0	1	0	26	0	0
	Blink	5	1	0	1	5	2
	No event	1	1	6	2	0	0

Table 8. Reliability statistics

	Up	**Down**	**Left**	**Right**	**Blink**	**Overall**
Sensitivity	92.0%	93.5%	100%	96.3%	35.7%	88.3%
Specificity	79.3%	90.6%	71.9%	86.7%	100%	82.8%

The algorithm had a different behaviour for eye movements and blinks. It had a very high sensitivity for the detection of movements in all directions (greater than 90%). No type of movement was missed, although some were wrongly classified (5% of them on average), and there were also false movement detections; so all in all the specificity of the algorithm was between 70% and 90%, depending on the direction of the movement.

However, the algorithm had a poor sensitivity to detect blinks (35.7%), which were often missed or misclassified as movements in the upward direction. However, the blink detection was extremely specific (100%), i.e. there was no false blink detection.

6 Sit-to-stand power

6.1 Objectives of the validation

The power exerted to raise the body during the sit-to-stand gesture (STS) is an important indicator of muscle strength, which can be used by My-AHA to assess the physical function of users. That variable is calculated as the product between the vertical force actuating at the body's centre of masses (CoM) and its vertical velocity. Those measures can be taken from force platforms or motion sensors (either optical or inertial), taking into account the following relations between the CoM vertical position (y_{CoM}), velocity (v_{CoM}), acceleration (a_{CoM}), and the vertical force actuating on the CoM (F_{CoM}), for a subject standing on ground with weight M:

$$F_{CoM} - 9.81\frac{m}{s^2} = M \cdot a_{CoM} = M \cdot \frac{dv_{CoM}}{dt} = M \cdot \frac{d^2 y_{CoM}}{dt^2}$$

The STS power is a transient signal that is often characterised by its peak value (Zijlstra et al. 2010), but to obtain a reliable measure of such peak value it is necessary to use accurate instruments and a high control the experimental conditions, in order to avoid the accumulation of errors in the derivation/integration and composition of the signals.

For unsupervised measures as the ones expected to be taken during the use of My-AHA, the effect of such errors can be minimised by taking a more robust parameter, like the average power during the rising phase. As observed in the left panel of Figure 13, there is an previous preparation phase in which the user pushes himself on the seat to take impulse and the power signal is negative, until the "seat-off" instant when the power starts to increase. After that moment, during the rising phase (marked with thick line) the power signal reaches its maximum and decays, until the body reaches its full height. The rising phase ends when the power signal approaches zero.

When the action is measured with a force platform (blue line in the left panel of the figure), the rising phase is delimited by the maximum value of the ground reaction force (the instant when the user is transferring the full weight of his body to the ground) and the point in which that force matches the user's weight, after decreasing and increasing again (Hirschfeld, Thorsteinsdottir, and Olsson 1999).

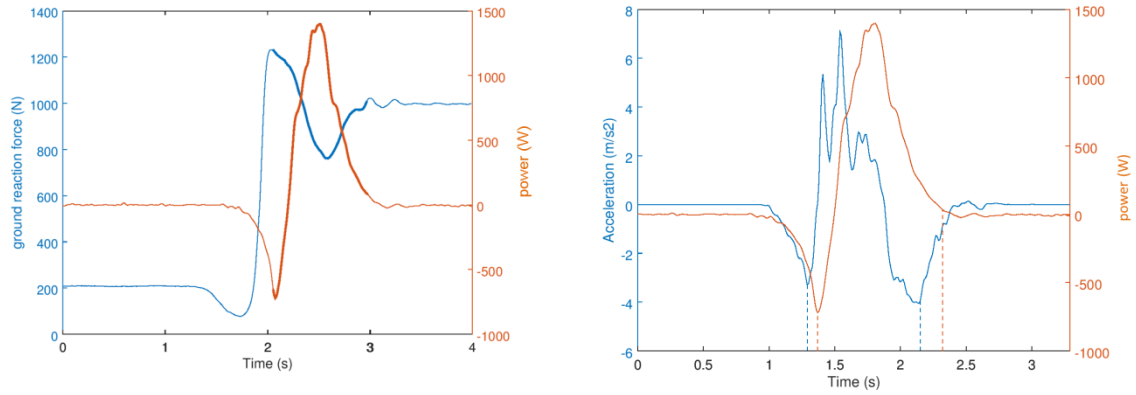

Figure 13. Power signal at the CoM compared with the reaction force measured at the ground (left) and the acceleration measured at the head. In the left panel the rising phase is marked with a thick line. In the right panel, the limits of the raising phase are compared with the local minima of the acceleration signal.

The dynamics of the body in the sit-to-stand gesture can be reduced to a simplified model such that the average power exerted during the rising phase is proportional to the ratio between the height gained by standing up and the time spent in the gesture. This model was analysed by Lindemann et al. (2003) with data from force plates. On the other hand, we propose using time landmarks of the acceleration signal measured by the inertial sensor (Figure 7 right), whose local minima occur a few instants before the onset and end of the rising phase.

6.2 Materials and methods

In a pilot experiment, two young adults (male, 100 kg, and female, 56 kg) repeated the STS gesture both at normal speed and as fast as possible — as replicating a timed sit-to-stand test —, such that 5 measures were taken at each speed. The experiments were conducted in the laboratories of IBV, and recorded simultaneously with synchronised dynamometric platforms and a photogrammetry system (Dinascan+Kinescan/IBV), and the inertial sensors integrated into the JINS MEME glasses. All devices were working at 100 Hz.

The average muscle power during the rising phase was calculated for each repetition with four different procedures:

- Full kinematic model: the acceleration and velocities of the body were calculated at each instant, (a_{CoM}, v_{CoM}), and the power signal was computed from them as by Zijlstra et al. (2010); the average power was calculated within the limits of the raising phase, obtained from the profile of the power signal as described in the previous section. The accelerations and velocities were estimated from sensors at the head's height, using two different measures:
    a) The position of a reflective marker measured by photogrammetry, differentiated two times to obtain its vertical acceleration and velocity.
    b) The acceleration measured by the JINS MEME sensor, integrated one time through the OFDRI procedure — optimised filtering and direct-reverse integration (Zok, Mazzà, and Della Croce 2004) to obtain an unbiased measure of the velocity.
- Simplified mode, as by Lindemann et al. . (2003): the height gained in the STS gesture was calculated as a fixed value from pre-computed standing and sitting heights, considering the anthropometry of the subjects, and the time of the raising phase was calculated from characteristic points of the measured signals. Those signals were:
    a) The profile of the ground reaction force measured by the platform, as in Lindemann's original article.
    b) The local minima of the acceleration signal measured by the JINS MEME sensor, around the its maximum value.

6.3 Results

The two variants of the full kinematic model were based on the same measures (acceleration and velocity), although they were obtained from different sources (the position of the marker measured by photogrammetry, and the acceleration measures by the inertial sensor). The gold standard in this case was the optical system, which was calibrated with an instrumental error for marker positions around 0.5 mm (Page et al. 2006). Figure 14 shows the acceleration, velocity and power signals obtained from each source for one of the measurements. It may be seen that they kinematic variables obtained from both sources were similar, although the measures derived from the acceleration signal have slightly smaller ranges, and the acceleration signal itself was less smooth. Therefore, the power values obtained with the inertial sensor were expected to be smaller than the ones obtained by photogrammetry.

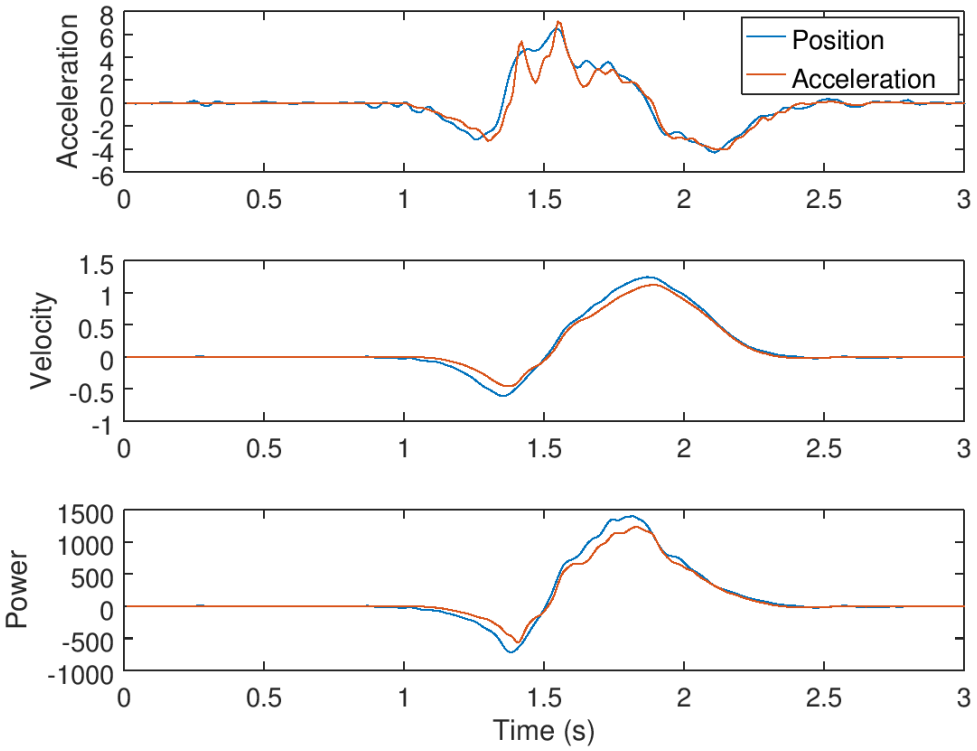

Figure 14. Kinematic variables (acceleration, velocity and power, in m/s^2, m/s and W, respectively) obtained by differentiation of positions measured by the optical system, and by integration of the inertial system.

The resulting estimates of the average power during the raising phase are compared between models and data sources in the combined Bland-Altman plot of Figure 15. The values obtained with the simple model and the data of the inertial sensor, based on the timing of characteristic points of the signal, are the easiest to implement in a system like My-AHA, and were used as reference, such that the values of all the other methods are compared against it.

The points of the plot are clearly grouped in four clusters, which correspond to the different combinations of subjects's weight (power values are proportional to weight, such that the values of the male, heavier subject are higher), and speed of the gesture within each subject (higher values for the fast STS gesture). Figure 16 shows a plot equivalent to that of Figure 15, but where the values are normalised by the user's weight, so that they are mainly clustered by the gesture speed.

That plot of normalised values shows more clearly some trends of the differences between the reference and the other methods. The differences observed with the full model with the photogrammetry data (Full-O), which can be deemed as the most precise estimates, were constrained in a range smaller than for the other methods, and showed a more random behaviour. The values obtained from the full model using the inertial data (Full-I) were systematically smaller, as expected from the comment on Figure 14. The values obtained with the simplified model and the force platform data (Simpl-F) were similar in most cases, but diverged in the trials of the heavier subject doing the fast gesture.

Even considering those differences, there was a good agreement among the measures, with an intra-class correlation coefficient (type 3) equal to 0.96 for normal speed, and 0.94 for the fast gestures.

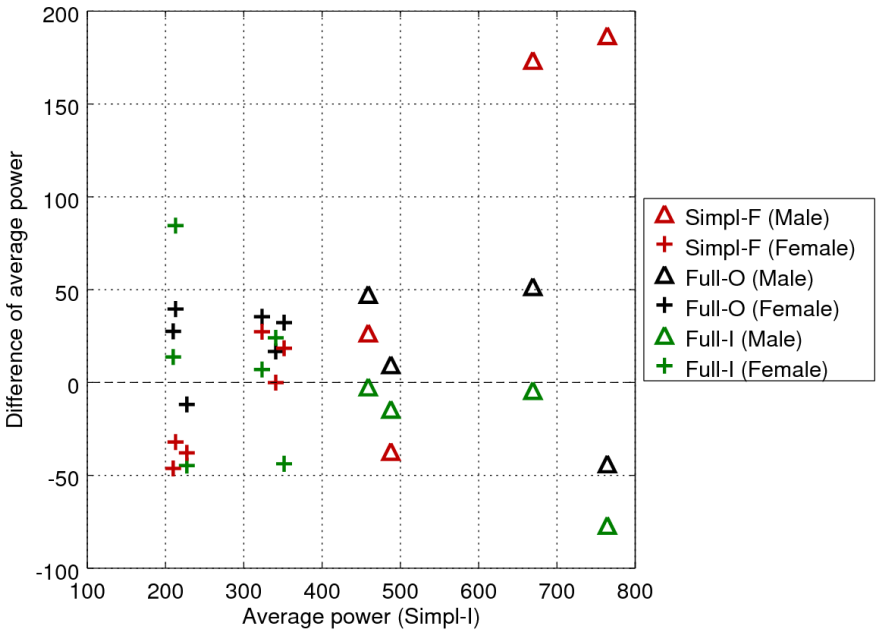

Figure 15. Difference plot of average STS power obtained with different methods. The reference value (in the X-axis) corresponds to the simplified model using the data of the inertial sensor (Simpl-I). The data points correspond to the simplified model using force data (Simpl-F), and the full model with data of the optical (Full-O) and inertial sensor (Full-I).

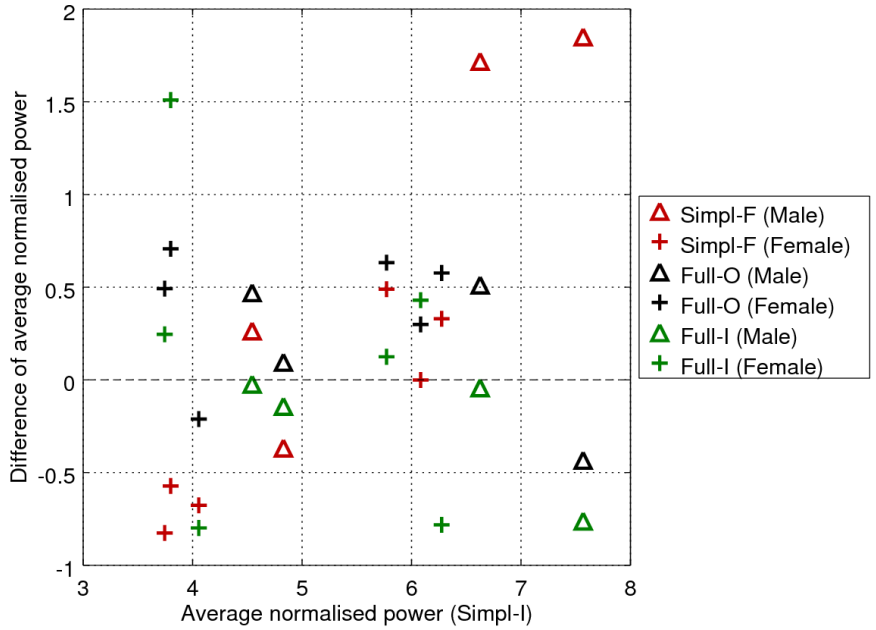

Figure 16. Difference plot of average STS power, normalised by the participant's weight.

7 Gait complexity

7.1 Objectives of the validation

The nonlinear measures of "complexity" are used as way of assessing the variability of gait signals, considering that the lack of periodicity of those signals is due to the nonlinear behaviour of neuromuscular dynamics, rather than to an amount of "randomness" added to an ideal periodic signal. Among the many parameters that can be used to characterise such complexity, we have chosen the measures of recurrence quantification analysis (RQA) and multiscale entropy (MSE) of the acceleration signal, which have been used in previous literature to evaluate gait in relation with the ageing process (Bisi and Stagni 2016; Riva, Bisi, and Stagni 2014).

As explained in D4.4, the biggest limitation of that kind of analysis is that it requires relatively long signals to produce reliable estimates of complexity measures, but on the other hand the computational resources consumed by such analysis grows following a power law of the time series length. Therefore, one of the objectives of the validation has been to verify that:

- it is possible to analyse sufficiently long fragments of acceleration data within reasonable limits of time and memory consumption for an application that should process thousands of daily data records; and
- real-life data records of older people produce fragments of the necessary length for such an analysis.

The other objective was to compare measures of RQA and MSE obtained from real-life data records with published values from previous studies.

7.2 Material and methods

7.2.1 Extraction of gait fragments

The minimum length of continuous gait periods that should be considered for the analysis was derived from published data of the number of strides required to obtain reliable estimates of gait complexity. According to the study published by Riva, Bisi, and Stagni (2014), RQA parameters and MSE (for scales equal or lower than 4) of the vertical acceleration signal are reliable for records that encopass more than 10 strides; and with 20 strides or more the same measures of anterior-posterior acceleration also have excellent reliability. Those reference values were compared with the statistics of walking activity recorded for 10 subjects during 1 month in April-May 2015 with Smart Companion. From that comparison we defined a minimum time period that should be considered to apply RQA and MSE measures.

With that minimum fragment length, we analysed continuous data of two older subjects (male 74 years, female 70 years) during 4 days, taken by DSHS in May 2016. Continuous gait periods over that length were obtained by threshold selection of the standard deviation and the short-time Fourier Transform (STFT) of the acceleration signal (Brajdic and Harle 2013), and we assessed the amount of continuous records that might be obtained with those data.

7.2.2 Measure of Multiscale Entropy

The original MSE algorithm published by Costa et al. (2002) has an important drawback when applied to uncontrolled measures, which is the risk of obtaining unstable or even undefined results. This risk is potentially higher in unassisted continuous measures, where there is less control of the signals. In order to reduce this risk, instead of the original MSE algorithm we computed the "Refined Composite Multiscale Entropy" (RCMSE), which has been demonstrated to be a robust estimate of MSE, and the "Modified Multiscale Entropy" (MMSE), which is even more robust and stable — but more expensive too (Humeau-Heurtier 2015). Their results were compared with the 4 days of data recorded by DSHS, to decide whether RCMSE would be a sufficiently robust and stable measure, or if it is necessary to resort to MMSE.

7.2.3 Assessment of RQA and MSE values

The RQA parameters selected for the analysis were: Recurrence Rate (RR), Determinism (DET), Average length of recurrent patterns (L), Divergence (DIV), Entropy (ENT), Laminarity (LAM) and Trapping Time (TT). MSE was analysed at scales between τ=1 and τ=8, as well as the "Complexity Index" (CI) obtained as cumulative sum of MSE values.

The values of RQA parameters and MSE estimates obtained from the 4 days of continuous, real-life data recorded by DSHS were compared with values published from previous studies in controlled settings (Bisi and Stagni 2016; Riva, Bisi, and Stagni 2014), to assess whether the results that might be obtained by continuous monitoring with devices connected to My-AHA might be representative of the results that are contained in laboratory studies that report differences in complexity measures related to the ageing process and gait anomalies.

In addition, we analysed the interquartile-median ratio (IMR) of the results within continuous gait periods of comparable length, in order to re-assess the reliability of those measures in a real-life setting. The IMR was the metric used by (Riva, Bisi, and Stagni 2014) to assess the repeatability of gait complexity measures in a controlled setting, considering that an IMR equal or smaller than 0.1 indicated an "excellent" repeatability.

7.3 Results

7.3.1 Extraction of gait fragments

Figure 17 shows that among older adults using the Smart Companion there was a relatively narrow band of gait cadences (running included), contained between 24 and 18 steps/minute. The average was 87.6 steps/min (std. dev. 6.6 steps/min).

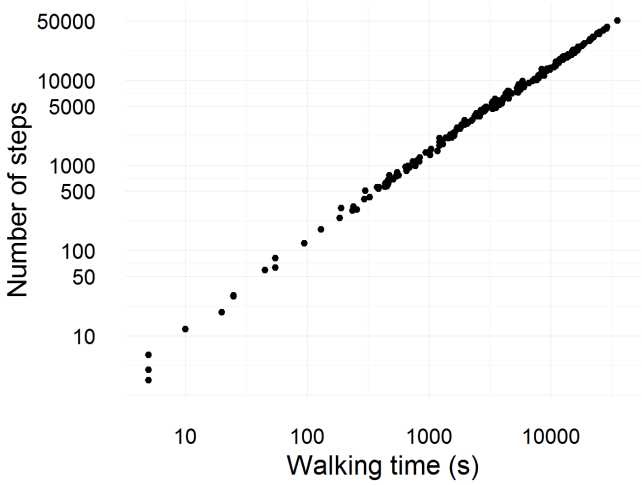

Figure 17. Relation between walking time and number of steps from Smart Companion data

Those data showed that gait fragments longer than 10 seconds may be expected to surpass the minimum threshold of 10 strides. Considering the slowest gait cadences, 40 seconds should suffice to obtain more than 20 steps. To fall in the safe side, we defined a minimum length of analysable gait fragments equal to 1 minute (60 seconds).

The continuous acceleration data recorded by DSHS during 4 days of two subjects contained 8 hours of walking activity distributed in 53 episodes longer than 1 minute. The average length of those episodes was 6.9 minutes (std. dev. 13.0 minutes), markedly biased towards shorter episodes (51% were 1-minute long).

Figure 18 shows the band of frequency spectra for such 1-minute gait fragments in the data from DSHS (in log-log scale). The average line and the limits of the bands are computed in a logarithmic scale (i.e. the average line is the geometric mean of the spectra measured at each fragment, and the limits correspond to the average multiplied or divided by a factor proportional to the standard deviation at each frequency point). It

can be seen that the dominant frequency is contained between 1 and 2 Hz, which corresponds to an average cadence similar but slightly faster than the reference data recorded by Smart Companion.

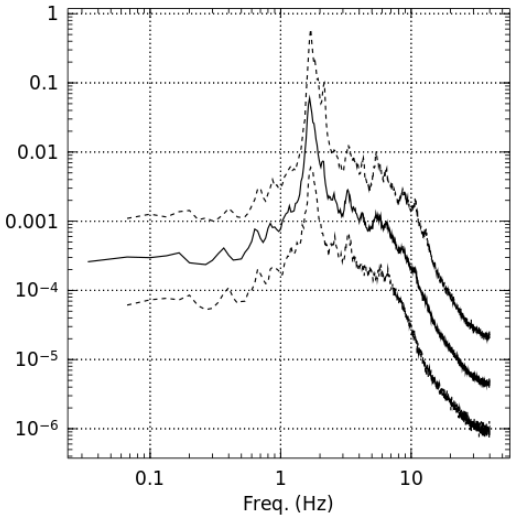

Figure 18. Band of frequency spectra of vertical acceleration during 1-minute gait fragments

The computation of RQA and MSE was implemented in Julia, a high-performance language for numerical computing (Edelman 2015). Processing each 1-minute fragment consumed 208 MB for RQA, 1 GB for RCMSE, and 7.5 GB for MMSE at eight scales; this took 0.6 s, 1.5 s and 20 s respectively, in a PC with Intel Core i3-4100M processor @ 2.5 GHz and 8 GB RAM, and Linux operating system (Ubuntu 16.04, Julia version 0.5.0). Memory and time consumption for RCMSE and MMSE grew with the number of scales involved, although with a converging rate for RCMSE, and diverging for MMSE (see Figure 19).

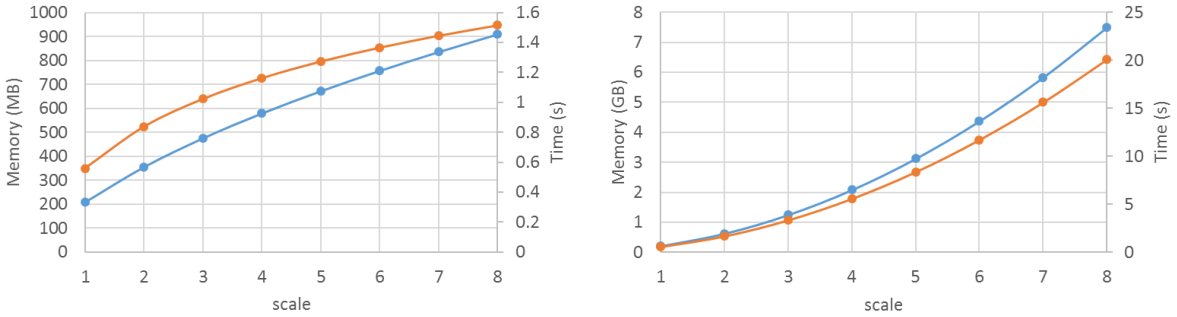

Figure 19. Memory and time consumption calculated for RCMSE (left) and MMSE (right)

7.3.2 Measure of Multiscale Entropy

Figure 20 shows the differences between RCMSE and MMSE for the extracted 1-minute fragments of vertical acceleration at different scales from τ=1 and τ=8. It may be observed that the dispersion of those differences grows proportionally to the size of the scale. Specifically, we observed a linear relation between the standard deviation of the difference ($\sigma_{RCMSE-MMSE}$) and the scale (τ), according to the equation:

$$\sigma_{RCMSE-MMSE} \sim k(\tau - 1)$$

That equation fitted the observed data with a factor $k = 4.2 \cdot 10^{-3}$ (standard error $4.3 \cdot 10^{-4}$, $R^2 = 0.93$).

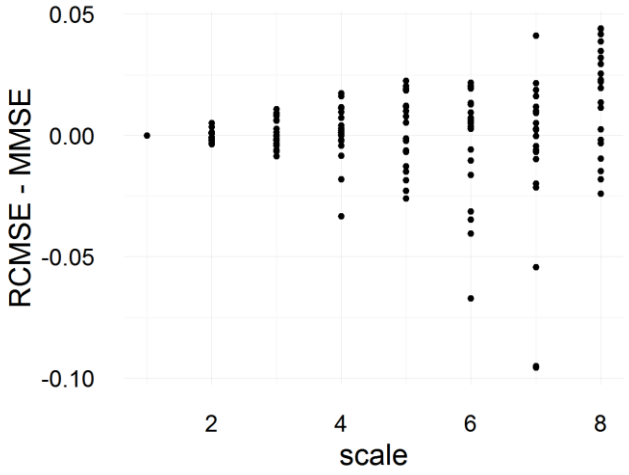

Figure 20. Difference between RCMSE and MMSE of vertical acceleration

In order to assess whether the difference between RCMSE and MMSE had any kind of bias (possibly depending on the scale, since at τ=1 there is no difference), a weighted linear model (accounting for the heteroscedasticity of the data) was fit to the differences, of the type:

$$RCMSE - MMSE \sim \beta_0 + \beta_1 \tau + \varepsilon \;\; : \;\; \varepsilon \sim N(0, k\tau)$$

The statistics of the fit (Table 9) showed that no parameter of the model was significantly different from zero. Therefore, there was no significant bias, and given the small size of the difference with respect to the expected values of either RCMSE or MMSE (see the next section), we can conclude that RCMSE is a sufficiently accurate and robust estimate of sample entropy, and may be preferred to MMSE, since it is several times faster (see the previous section).

Table 9. Statistics of the linear model fitted to the RCMSE-MMSE difference

	Estimate	Std. error	t(138)	p-value
Intercept (β_0)	$-1.08 \cdot 10^{-3}$	$2.24 \cdot 10^{-3}$	-0.483	0.630
Slope (β_1)	$0.39 \cdot 10^{-4}$	$6.15 \cdot 10^{-4}$	0.635	0.527

7.3.3 Assessment of RQA and MSE values

The observed values of RQA parameters and the selected measures of MSE (RCMSE and the complexity index, CI), are summarised in Table 10 and Table 11. The curves of RCMSE as a function of the scale are also plotted in Figure 21 for greater clarity.

Table 10. Summary of RQA parameters for 1-minute gait periods

	Vertical		Anterior-posterior	
	Mean	**(Std. dev.)**	**Mean**	**(Std. dev.)**
RR (%)	2.66	(1.11)	5.24	(6.77)
DET (%)	83.55	(6.76)	75.32	(9.95)
L (ms*)	124.20	(24.46)	105.72	(46.63)
DIV ($\times 10^{-3}$)	1.16	(5.45)	13.26	(4.75)
ENT	1.93	(0.22)	1.69	(0.39)
LAM (%)	81.16	(10.86)	83.30	(7.55)
TT (ms*)	74.42	(21.04)	105.17	(67.34)

(*) L and TT are usually expressed as a number of data points. They are transformed into time units to make them comparable with results of measures taken at different sampling rates.

Table 11. Summary of RCMSE and CI for 1-minute gait periods

	RCMSE Vertical		RCMSE AP		CI Vertical		CI AP	
	Mean	**(Std. dev.)**	**Mean**	**(Std. dev.)**	**Mean**	**(Std. dev.)**	**Mean**	**(Std. dev.)**
τ=1	0.499	(0.049)	0.594	(0.173)	0.499	(0.049)	0.593	(0.173)
τ=2	0.658	(0.088)	0.945	(0.279)	1.157	(0.132)	1.539	(0.451)
τ=3	0.784	(0.138)	1.173	(0.320)	1.941	(0.263)	2.713	(0.769)
τ=4	0.869	(0.179)	1.291	(0.329)	2.810	(0.430)	4.004	(1.092)
τ=5	0.907	(0.199)	1.347	(0.342)	3.717	(0.614)	5.350	(1.427)
τ=6	0.927	(0.199)	1.368	(0.366)	4.643	(0.802)	6.715	(1.785)
τ=7	0.917	(0.192)	1.371	(0.382)	5.560	(0.985)	8.090	(2.160)
τ=8	0.873	(0.183)	1.371	(0.368)	6.433	(1.157)	9.461	(2.520)

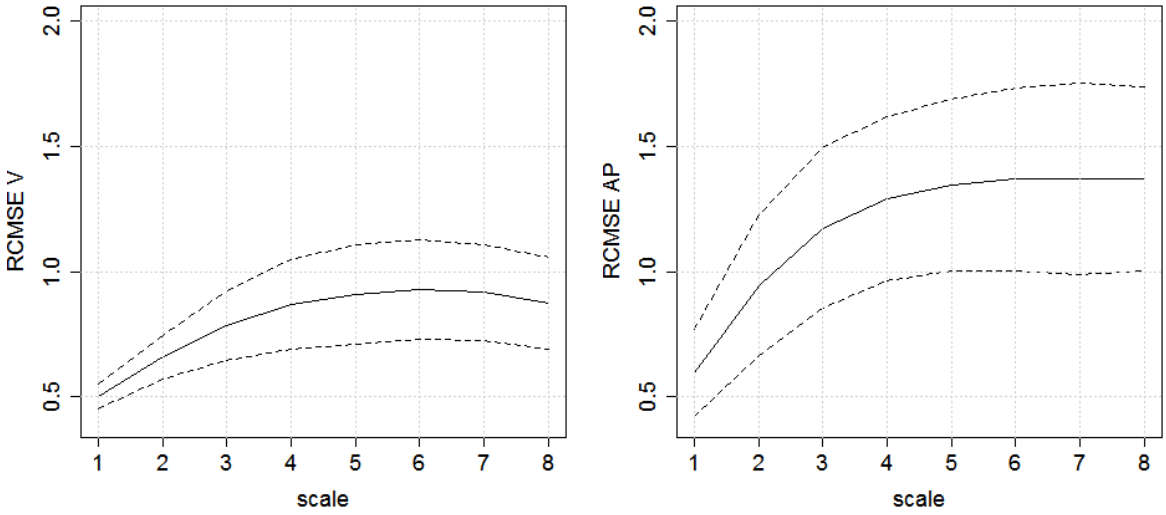

Figure 21. RCMSE mean curves (± std. dev.) of vertical (left) and anterior-posterior (right) acceleration

The study with the greatest amount of numeric details about RQA and MSE measures of gait dynamics was published by Riva et al. (2014), who reported greater RR (between 10% and 15%, between 3 and 5 times greater than the ones observed by us), but DET and L values that agreed with our results (considering the differences in sampling rates between studies).

The RCMSE values observed by us also coincided with previously reported values of MSE (only given for scales equal or smaller than 4 or 6) in the case of vertical acceleration, but we obtained greater values in the anterior-posterior direction.

The distributions of IMR are shown in Figure 22 (for RQA parameters) and Figure 23 (for RCMSE and CI). "Excellent" values (IMR < 10%) were always obtained for DET, ENT and LAM in both directions (vertical and anterior-posterior), and also for TT of vertical acceleration. Setting a reference value of IMR < 15%, we can also consider L in both directions, as well as RCMSE and CI values for scales $\tau \leq 4$ (let aside outliers).

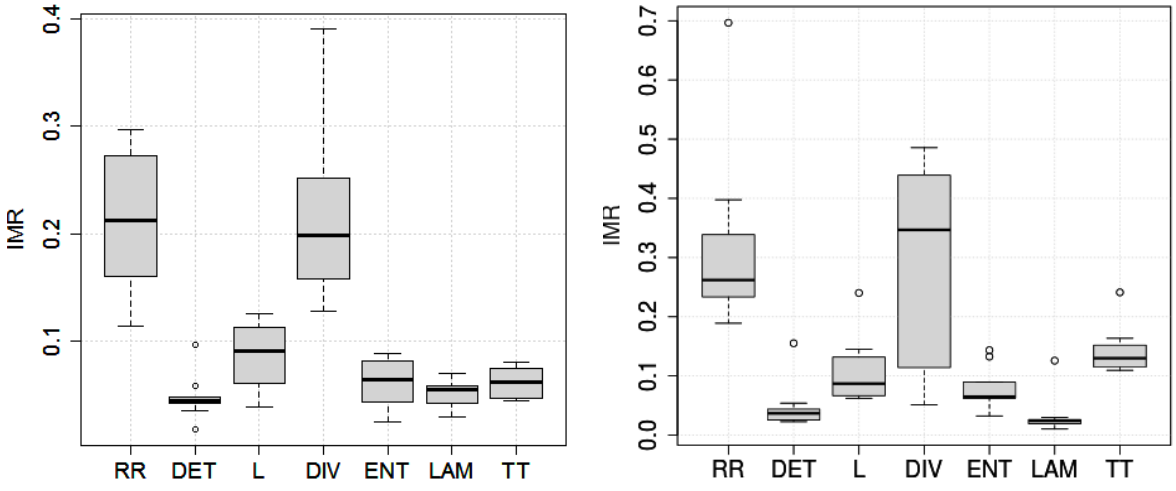

Figure 22. IMR of RQA parameters for vertical (left) and anterior-posterior (right) acceleration

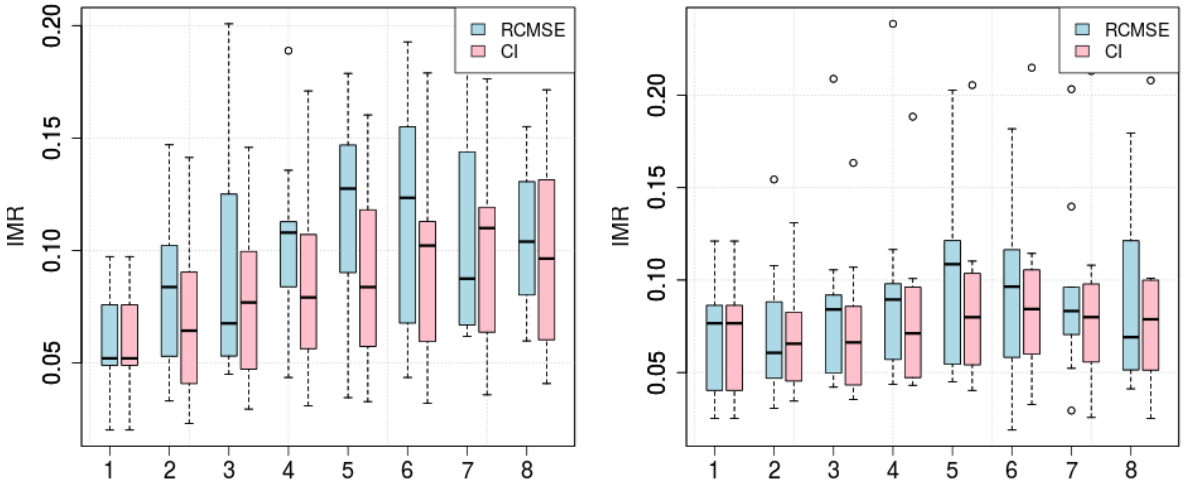

Figure 23. IMR of RCMSE and CI of vertical (left) and anterior-posterior (right) acceleration

8 Conclusions

The six algorithms selected for validation have been tested in pilot experiments in controlled conditions to check their performance, and yielded the following results:

- **The analysis of HRV** with the Mio wristband resulted to be reliable for long-term and low-frequency features (SDNN and VLF, respectively). On the other hand short-term and high frequency features, related to quick changes in the beat-to-beat time intervals, could not be successfully retrieved from the wristband, regardless of the type of activity performed by the subjects (quiet or walking/running). This was mainly attributable to the filters implemented in the acquisition software in order to reduce the effects of noise and movement-induced artefacts, since the sensor itself is of high-quality profile compared to similar products in the market. The short term/high frequency information of the NN signal is just removed from the signal delivered by the sensor, so it is not possible to obtain those features.
- **Speech analysis** turned out to be very reliable to evaluate the fatigue of the subjects, compared with the scores of the Karolinska Sleepiness Scale, even considering small changes in the state of the subjects. The KSS scores predicted by the speech analysis were consistent with the self-reported values using the standard questionnaires, but presented a smaller inter-subjects variability. In other words, the analysis of speech features produced estimates of the subjects' fatigue that were less dependent on the difference between subjects than relying on their self report. The differences between the self-reported and predicted KSS scores were random, homogeneous and unbiased errors of the same order of magnitude as the variability within subjects.
- **The activity recognition algorithm with EOG data** from the JINS MEME glasses could discriminate between reading, watching TV and drinking, with a success rate around 86%, much better than popular basic classifiers. The outcomes differed depending on the training data and codebooks, and it was detected that there may be problems with noise and artefacts produced by manipulation of the glasses.
- **The detection of eye movements and blinks** with EOG data was also successful, with very high sensitivities and good specificities for the recognition of directional movements. On the other hand, the sensitivity to detect blinks is smaller, since they are faster actions that may be missed due to the limited sampling rate of the glasses (even in the academic version), and show EOG features similar to the vertical movement of the eyes in the upper direction. As happens with the algorithm for activity recognition, the detrimental effect of motion artefacts should be considered. (There are other methods aimed at reducing those artefacts, presented in D4.3).
- **The analysis of sit-to-stand power** yielded promising results, even though measuring the power signal is a complex challenge, and different methods with high-quality instruments that could be considered "gold standards" may provide different results. A simple approach that might be implemented in My-AHA, based on timing features of the acceleration signal, provided values of the average power during the rising phase that were in the same range as the gold standard measured with photogrammetry.
- **The measure of gait complexity** proved to be feasible in terms of time and computer resources expenditure, in spite of the big amounts of data that have to be processed, when gait episodes are analysed in one-minute intervals. It has been demonstrated that RQA and MSE measures from real-life data are repeatable when analysed in such intervals, and are also consistent with previous results reported in literature.

The pilot experiments reported in this deliverable have some limitations: in most cases have been conducted with small subject samples, except for the speech analysis and the activity recognition, which required big data bases to train the machine learning algorithms, and except for gait analysis, the participants in the pilots were young adults. The differences in the measures and behaviours between young and older adults must be taken into account for further developments in My-AHA. The validation of the speech recognition algorithm was also limited to German-speaking subjects.

These results, together with other technical aspects that have been evaluated in technical discussions, will be analysed by My-AHA's Consortium to make future decisions regarding the development of the platform, and the sensors and signal pre-processing modules that will be connected to it.

References

Antelmi, Ivana, Rogério Silva De Paula, Alexandre R. Shinzato, Clóvis Araújo Peres, Alfredo José Mansur, and Cesar José Grupi. 2004. "Influence of Age, Gender, Body Mass Index, and Functional Capacity on Heart Rate Variability in a Cohort of Subjects without Heart Disease." *The American Journal of Cardiology* 93 (3): 381–85. doi:10.1016/j.amjcard.2003.09.065.

Bisi, M. C., and R. Stagni. 2016. "Complexity of Human Gait Pattern at Different Ages Assessed Using Multiscale Entropy: From Development to Decline." *Gait & Posture* 47 (June): 37–42. doi:10.1016/j.gaitpost.2016.04.001.

Brajdic, Agata, and Robert Harle. 2013. "Walk Detection and Step Counting on Unconstrained Smartphones." In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 225–234. UbiComp '13. New York, NY, USA: ACM. doi:10.1145/2493432.2493449.

Chan, Hsiao-Lung, Ming-An Lin, Pei-Kuang Chao, and Chun-Hsien Lin. 2007. "Correlates of the Shift in Heart Rate Variability with Postures and Walking by Time-Frequency Analysis." *Computer Methods and Programs in Biomedicine* 86 (2): 124–30. doi:10.1016/j.cmpb.2007.02.003.

Costa, Madalena, Ary L. Goldberger, and C.-K. Peng. 2002. "Multiscale Entropy Analysis of Complex Physiologic Time Series." *Physical Review Letters* 89 (6): 068102. doi:10.1103/PhysRevLett.89.068102.

Edelman, A. 2015. "Julia: A Fresh Approach to Parallel Programming." In *Parallel and Distributed Processing Symposium (IPDPS), 2015 IEEE International*, 517–517. doi:10.1109/IPDPS.2015.122.

Hirschfeld, Helga, Maria Thorsteinsdottir, and Elisabeth Olsson. 1999. "Coordinated Ground Forces Exerted by Buttocks and Feet Are Adequately Programmed for Weight Transfer During Sit-to-Stand." *Journal of Neurophysiology* 82 (6): 3021–29.

Humeau-Heurtier, Anne. 2015. "The Multiscale Entropy Algorithm and Its Variants: A Review." *Entropy* 17 (5): 3110–23. doi:10.3390/e17053110.

Kaida, Kosuke, Masaya Takahashi, Torbjörn Åkerstedt, Akinori Nakata, Yasumasa Otsuka, Takashi Haratani, and Kenji Fukasawa. 2006. "Validation of the Karolinska Sleepiness Scale against Performance and EEG Variables." *Clinical Neurophysiology* 117 (7): 1574–81. doi:10.1016/j.clinph.2006.03.011.

Lindemann, Ulrich, Holger Claus, Michael Stuber, Peter Augat, Rainer Muche, Thorsten Nikolaus, and Clemens Becker. 2003. "Measuring Power during the Sit-to-Stand Transfer." *European Journal of Applied Physiology* 89 (5): 466–70. doi:10.1007/s00421-003-0837-z.

Lipsitz, L. A., J. Mietus, G. B. Moody, and A. L. Goldberger. 1990. "Spectral Characteristics of Heart Rate Variability before and during Postural Tilt. Relations to Aging and Risk of Syncope." *Circulation* 81 (6): 1803–10. doi:10.1161/01.CIR.81.6.1803.

Nunan, David, Djordje G. Jakovljevic, Gay Donovan, Lynette D. Hodges, Gavin R. H. Sandercock, and David A. Brodie. 2008. "Levels of Agreement for RR Intervals and Short-Term Heart Rate Variability Obtained from the Polar S810 and an Alternative System." *European Journal of Applied Physiology* 103 (5): 529–37. doi:10.1007/s00421-008-0742-6.

Page, A., H. Rosario, V. Mata, J. V. Hoyos, and R. Porcar. 2006. "Effect of Marker Cluster Design on the Accuracy of Human Movement Analysis Using Stereophotogrammetry." *Medical & Biological Engineering & Computing* 44 (12): 1113–19. doi:10.1007/s11517-006-0124-3.

Pereira, Marta LG Freitas, Marina von Zuben A Camargo, Ivan Aprahamian, and Orestes V Forlenza. 2014. "Eye Movement Analysis and Cognitive Processing: Detecting Indicators of Conversion to Alzheimer's Disease." *Neuropsychiatric Disease and Treatment* 10 (July): 1273–85. doi:10.2147/NDT.S55371.

Riva, F., M. C. Bisi, and R. Stagni. 2014. "Gait Variability and Stability Measures: Minimum Number of Strides and within-Session Reliability." *Computers in Biology and Medicine* 50 (July): 9–13. doi:10.1016/j.compbiomed.2014.04.001.

Vanderlei, L. C. M., R. A. Silva, C. M. Pastre, F. M. Azevedo, and M. F. Godoy. 2008. "Comparison of the Polar S810i Monitor and the ECG for the Analysis of Heart Rate Variability in the Time and Frequency Domains." *Brazilian Journal of Medical and Biological Research* 41 (10): 854–59. doi:10.1590/S0100-879X2008005000039.

Weippert, Matthias, Mohit Kumar, Steffi Kreuzfeld, Dagmar Arndt, Annika Rieger, and Regina Stoll. 2010. "Comparison of Three Mobile Devices for Measuring R–R Intervals and Heart Rate Variability:

Polar S810i, Suunto t6 and an Ambulatory ECG System." *European Journal of Applied Physiology* 109 (4): 779–86. doi:10.1007/s00421-010-1415-9.

Zijlstra, Wiebren, Robertus Wilhelmus Bisseling, Stephan Schlumbohm, and Heribert Baldus. 2010. "A Body-Fixed-Sensor-Based Analysis of Power during Sit-to-Stand Movements." *Gait & Posture* 31 (2): 272–78. doi:10.1016/j.gaitpost.2009.11.003.

Zok, Mounir, Claudia Mazzà, and Ugo Della Croce. 2004. "Total Body Centre of Mass Displacement Estimated Using Ground Reactions during Transitory Motor Tasks: Application to Step Ascent." *Medical Engineering & Physics*, This issue contains a special section on Neuromodelling, 26 (9): 791–98. doi:10.1016/j.medengphy.2004.07.005.

