Essential Feature Extraction of Photoplethysmography Signal of Men and Women in Their 20s

Yasser Abd Djawad*, Andi Mu’nisa, Pangayoman Rusung, Abdi Kurniawan, Irma Suryani Idris, and Mushawwir Taiyeb

Universitas Negeri Makassar, JL. AP Pettarani, Makassar 90222, Indonesia
*E-mail: yasser.djawad@unm.ac.id (Corresponding author)

Abstract. This study aims to extract the essential features of Photoplethysmography (PPG) signal of men and women in healthy subjects using Power Spectral Density (PSD) and Detrended Fluctuation Analysis (DFA). A PPG instrument was used to obtain the PPG signal of 15 men and 15 women. Using PSD, four frequency bands were selected to divide the spectral component. The areas within the frequency bands relative to the total area were computed as features of the signals. Furthermore, using DFA, the average fluctuation F(w) was computed. The feature extraction using this technique produced 4 features from different windows. Hurst exponent was calculated to analyse the characteristics of the time series. For comparing the feature extraction techniques, Heart Rate (HR) and Peak to Peak Interval (PPI) were computed. Additionally, F and T tests for all techniques were computed to determine the differences between man and woman features that have been gathered using these two techniques. The results indicate that the features of PPG signals of men and women using PSD and DFA were significantly different. In order to evaluate the results, a clustering analysis was applied to the results using K-means clustering technique. The clustering plots show that the features were well distributed into the two groups.

Keywords: Feature extraction, frequency bands, spectral component, average fluctuation.
1. Introduction

The measurement of blood volume is normally conducted in vivo using a technique called chamber-plethysmography. In order to conduct the experiment, the technique uses a chamber which makes the measurement need a plenty of space and complex measurement. However, in some cases, a simple technique that provides an immediate and advancement analysis is required. A non-invasive technique is used to obtain the blood volume changes called photoplethysmography (PPG). A PPG measurement device consists of an infra-red light source and a photo-detector. The infra-red light is illuminated through the skin. The penetration depth of light depends on the absorption and scattering coefficients of tissue. Therefore, the use of an infra-red light is preferred since it is more stable over time compared to red light [1]. The PPG instrument can be used in two modes, reflection mode and transmission mode. In the reflection mode, the light source and the photo-detector are placed side by side. The infra-red light is illuminated to the skin and the reflection light from tissue and bones is accepted by a photo-detector. Meanwhile in the transmission mode, the light is illuminated through the skin and detected by a photo-detector, which means that the light source and the photo-detector are facing each other. The light intensity accepted by the photo-detector represents the blood volume and the heart rate [2].

The PPG signal has a fundamental frequency of around 1 Hz. However, it has unequal periods. The PPG signal consists of pulsatile waveforms and DC components. The pulsatile waveforms related to the heart rate and the DC component represent the tissues and the average of blood volume [2]. A considerable number of studies on the processing of PPG signals using statistical and theory of non-linear dynamical analyses have been carried out [3, 4, 5, 6]. In addition, several time series analyses have increasingly been introduced to analyse the PPG signal, including fractal analysis. The fractal analysis provides an ability to measure the pattern behaviour of the time series and serial correlation over time windows [7]. Some of those fractal analyses introduce a Hurst exponent $H$ which indicates the level of autocorrelated properties of the time series. One of these techniques is Detrended Fluctuation Analysis (DFA) [8].

There are increasing concerns about the differences of internal organ characteristics of men and women. Multiple factors have been cited by researchers to obtain the gender related differences of internal organs. [9] found that Body Mass Index (BMI) of men is higher than that of women. Using different approaches, numerous studies have attempted to explain the differences of the cardiovascular system of men and women, including Heart Rate Variability (HRV) or Heart Rate (HR). These previous studies have reported the relationships between the HRV or HR of ECG signals with the gender-related differences. All research findings indicate that there are significant differences of HRV parameters between men and women. [10] define that the high frequency heart rate fluctuation of men is greater than that of women in young subjects, and the heart rate dynamics of women is more complex than that of men. [11] also found that men in younger group (29-46 years old) had higher HRV parameters compared to women. Meanwhile there were no significant differences in the older group (64-74 years old). [12, 13] observed considerable differences between HRV parameters of men and women where the HRV parameters are higher in men. However, the features of PPG signal of men and women are still under explored. They have not been much investigated by researchers.

The HRV analysis is the most common analysis of cardiovascular system. PPG-HRV analysis was introduced as an alternative to analyse the PPG signal. [14, 15] applied HRV analysis to the PPG signal after pre-processing the PPG signal. Two features of HRV parameters (SDNN and rMSSD) could be extracted using this technique. However, according to the ECG-HRV analysis, only SDNN parameter value has a significant difference between men and women while there is no difference of rMSSD between men and women [12, 13]. Further research has reported that PPG-HRV technique generates unsatisfying results [16]. Previous studies have revealed the presence of fractal components in PPG signal [17, 18]. In particular, [19] applied DFA to PPG signal to study the hemodialysis of patients with diabetes mellitus by analysing the change of
scaling exponent $\alpha$. Fractal signal analysis has also been applied to a number of bio-signals to obtain important information, such as [20] who investigated the number of fractal component in human EEG signals using Coarse Graining Spectral Analysis (GCSA). [21] analysed EEG signals of 19 outpatients during hypnotising period using DFA. The use of DFA to discriminate EEG, EOG, EMG and ECG signals of sleep stages and sleep apnea severity was demonstrated by [22]. Therefore, this study focuses on extracting more features of the PPG signal of the gender-related differences using PSD and DFA.

2. Methods

2.1. Instrumentation

The PPG measurement was conducted using an infra-red sensor (HRM-215E) connected series with two stages of filter and amplifier. The sensor was connected to a High Pass Filter (HPF) with frequency cut off ($f_c$) of 0.5 Hz to attenuate the DC component from the sensor and to avoid the PPG signal from going to the saturation mode. The output of HPF was connected to the first stage of active filter which consists of an op-amp based active Low Pass Filter (LPF) with frequency cut off of 3.4 Hz and gain ($A$) of 48. In order to pull down the swing output from the first stage, a similar HPF was placed before the second stage. The second stage filter and amplifier were connected to this HPF. Thus, the overall gain was approximately 2300. The last stage was an op-amp based buffer with a unity gain to match the impedance to the following instrument. Analog output of the PPG measurement device was connected to Arduino Uno for data acquisition with a transfer rate of 9600 bit per second. Arduino uno was connected to a laptop for saving the data as a csv (comma separate values) file for analysis using R language.
Fig. 2. Peak to Peak Interval is the width between the peaks of PPG signal.

2.2. Subjects

The data acquisition involved two experimental groups of healthy man and woman volunteers who were between 20 and 25 years old since they provide strong emotional strain and less disease [23, 24]. Each group consists of 15 volunteers. During data acquisition, the two groups were seated comfortably and their left hand was bent down on a table with an angle of 130° ± 5° in a room with a temperature of 26-28°C. The data were recorded for 5 minutes as suggested by [25] for data processing using frequency domain method to the tip of the subject’s left finger. Before the PPG pulse waves were acquired, the systolic and diastolic arterial blood pressure of the volunteers were taken. Subjects who smoked, had their period and took medications were excluded from this data acquisition. In this study, screening blood tests were not conducted. The procedure of data acquisition had been approved by the local Ethics Committee.

2.3. Heart Rate and Peak to Peak Interval

Heart rate (HR) is the number of heartbeat in a minute when contractions occur. The HR of adults when taking a rest is 60 to 100 beats per minute (bpm). Some factors that may affect HR are body position, body size and medication used. Heart rate is calculated using this formula:

\[
bpm = \frac{\text{Sampling rate}}{\text{PPI}(i + 1) - \text{PPI}(i)} \times 60; \quad (1)
\]

Peak to Peak Interval (PPI) is the width between two peaks as shown in Figure 2. The PPI represents cardiac beat-to-beat interval of PPG signal.

2.4. Power Spectral Density

In this study frequency domain analysis was conducted because in order to compute the parameters, it only needs short-time recording compared to time domain method [25]. To compute the frequency domain parameters, Power Spectral Density (PSD) analysis was used. PSD reveals the distribution of power signal over frequency, which can be defined as:

\[
S_{xx} = \lim_{T \to \infty} \mathbb{E}[|\hat{x}_T(i)|^2] \quad (2)
\]

where \(\hat{x}_T(i)\) is Fourier transform of signal \(x(t)\).

2.5. Detrended Fluctuation Analysis

DFA is a statistical technique for scaling long range correlations in a time series [8]. Consider, a 1-D time series \(X(i), i=1,...,N\), where \(N\) is the length of the time series. The integrated time series can be computed using the following formula:
The method of DFA with a width of window. Similar method repeats for various width of window \( w \).

\[
y(k) = \sum_{i=1}^{N} (X(i) - X_{avg})
\]

where \( X_{avg} \) is the mean value of the time series \( X(i) \). The time series is divided into several windows (or segments) with width \( w \). To determine the width of the windows \( w \), the smallest value must be chosen to begin and the average fluctuation \( F(w) \) of the signal is computed according to :

\[
F(w) = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (y(k) - y_n(k))^2}
\]

The process is repeated for all considered \( w \) and it ends on the largest desirable value like \( N/2 \) [7]. In each window, a double logarithmic graph, \( \log F(w) \) vs \( \log w \), is created to retrieve best approximation of data trend or linear approximation of the data between these two parameters by applying standard least square regression as shown in Figure 3. Hurst exponent (H) is a way to identify the autocorrelation in a time series. H values are divided by three categories. A value of 0.5 indicates an uncorrelated data or that the time series does not have previous value (no memory). Meanwhile, values of \( 0 < H < 0.5 \) indicate that time series has a negative autocorrelation and values of \( 0.5 < H < 1 \) indicate a positive autocorrelation of time series. The value of Hurst exponent describes the power-law correlation and the smoothness of time series [8].

### 2.6. F-Test and T-Test

In order to determine the difference between the features of men and the feature of women, F-test and T-test (two tailed) were performed for each window. The tests were performed to analyse the difference of men and women features, from PSD and DFA computation. The following formula is used for T-test :

\[
T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}}
\]

where \( \bar{Y}_1 \) and \( \bar{Y}_2 \) are the sample means, \( s_1 \) and \( s_2 \) are the standard deviations of the samples and \( N_1 \) and \( N_2 \) are the sample sizes. The results of p-value from F-test indicates whether the variances of the two groups of data are equal or not. When the p-value > 0.05, it is assumed that the variances are equal and if the p-value < 0.05 then it is assumed that the variances are not equal. Based on the F-test, the T-test is conducted. When p-value of T-test < 0.05 then it is assumed that the data are different.
2.7. K-means Clustering

Clustering is a process to collect data that have similar characteristics into sets of groups. One of the clustering techniques is K-means clustering. K-means algorithm puts the data point in a dimensional space into K clusters. Each cluster is parameterized by its means ($m^{(k)}$). K-means clustering algorithm is described as follows:

Algorithm 1 K-means algorithm

1: Choose K means $m^k$ that represent centroids
2: repeat
3: assign all data points to the closest centroids in the form K clusters

$$k^{(n)} = \arg\min_x d(m^{(k)}, x^{(n)})$$ (6)

4: when all data points have been assigned, recalculate the centroid of each cluster
5: until the centroids do not longer move

3. Results

Baseline characteristics in Table 1 reveal that there were notable significant differences between men and women in terms of systolic blood variable. The average of systolic blood pressure of men was 116.7 while the average blood pressure of women was 108. The difference is shown in the average of diastolic blood pressure, where the average of diastolic blood pressure for men was 83.3 and the average diastolic blood pressure for women was 77.3. This results align with previous research [10, 26, 27] which reported that women have lower blood pressure of about 5 to 10 mm Hg compared to men.

Table 1. Baseline characteristics of the subjects.

<table>
<thead>
<tr>
<th>Variable</th>
<th>men</th>
<th>women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ages</td>
<td>22 ± 1.5</td>
<td>23 ± 1.3</td>
</tr>
<tr>
<td>Systolic blood pressure (mm Hg)</td>
<td>116.7 ± 0.8</td>
<td>108 ± 0.5</td>
</tr>
<tr>
<td>Diastolic blood pressure (mm Hg)</td>
<td>83.3 ± 0.7</td>
<td>77.3 ± 0.5</td>
</tr>
</tbody>
</table>

3.1. Heart Rate and Peak to Peak Interval

Fig. 4. Box and whiskers plot of HR and PPI of volunteers.

Figure 4 shows Box and Whiskers plot of HR and PPI of PPG signal of volunteers which are standard techniques for PPG features extraction. The plot shows that there was no significant
Table 2. Probability of HR and PPI features.

<table>
<thead>
<tr>
<th>Test</th>
<th>Heart rate</th>
<th>Peak to Peak Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>p of F test</td>
<td>0.3562</td>
<td>0.0053</td>
</tr>
<tr>
<td>p of T test</td>
<td>0.3142</td>
<td>0.0032</td>
</tr>
</tbody>
</table>

Table 3. Probability of features extraction from the PSD analysis.

<table>
<thead>
<tr>
<th>Test</th>
<th>Frequency bands</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>p of F test</td>
<td>1.25E-13</td>
</tr>
<tr>
<td>p of T test</td>
<td>0.006070</td>
</tr>
</tbody>
</table>

difference between the HR of men and women. However, there was a significant difference between PPI features of men and women.

F and T Test of HRV and PPI are shown in Table 2. The value of p of F test of HR was > 0.005 which means that the variances were equal. The p value of T test of HR was also > 0.05 that means the HR features of men and women were not different. Different results were shown by PPI features in which p value of PPI features is < 0.005 which reveals that the PPI values between men and women were different.

3.2. Power Spectral Density

The technique divides spectral components into four ranges of frequency bands which are different from HRV parameters. Two samples of men and women were taken for ranges frequency band dividing analysis. The frequency band ranges have been chosen based on the analysis which was conducted by comparing two samples of PSD of men and women as shown in Figure 5. The spectral component ranges are Frequency Band 1 (0.5-0.8 Hz), Frequency Band 2 (0.8-1.3 Hz), Frequency Band 3 (1.3-1.8 Hz) and Frequency Band 4 (1.8-2.5 Hz). Frequency Band 1 was chosen since most of men spectral components are within this band. Meanwhile, most of women’s spectral components were within Frequency Band 2. Frequency Band 3 was selected because within this band most of men’s spectral components show higher magnitude compared to women. Finally, Frequency Band 4 was chosen to provide an additional band even though men’s and women’s spectral component within this band showing no significant differences.

The four areas within the frequency bands were computed. In order to obtain the features of the signals, the area relative (RA) to the total area was calculated as follows:

\[ RA = \left( \frac{AFB}{TA} \right) \times 100; \]  

where AFB is the area within the frequency band and TA is the total area within all frequency bands. The box and whiskers graph of the relative area is shown in Figure 6.

In order to obtain the difference of the result from PSD analysis, the F and T test (two tailed) were conducted and the result is shown in Table 3. It shows that variances of relative area within Frequency Band 1, Band 2 and Band 3 were unequal since p-values of F test < 0.05. Meanwhile, Frequency Band 4 had an equal variance. However, the T test showed that only Frequency Band 1 and Frequency Band 2 provided significant a difference where p-values < 0.05 and there was no significant difference for Frequency Band 3 and Band 4.

3.3. Detrended Fluctuation Analysis

The data acquisition was performed to each subject which produced 15 PPG time series signal for men and 15 PPG time series signal for women. Each time series of the signal was analysed.
Fig. 5. Four ranges from two samples of PSD analysis of: (a) men; (b) women.

Fig. 6. Average of relative area of 4 frequency bands of men (M) and women (W) using DFA. The DFA computation created 4 windows. The windows were multiple of two from 4 to 32 as shown in Figures 7(a) and 7(b).

Figure 8 shows the result of DFA computation of men (M) and women (W) for each window. The graphs show that for all windows, F(w) of women were higher than that of men. There were significant differences between the data of women and the data of men as the window increases. However, windows above 32 in this study were not considered because they generated almost similar features. Meanwhile, Hurst exponent result shows that the values were located between 0.5 and 1, which indicates that the DFA features of PPG signals of men and women had a memory or had a positive autocorrelation, as shown in Figure 9.

Table 4 shows the p-value of F-test and T-test. The results of F-test show that the p-values on width of window 4,8,16 and 32 were bigger than 0.05 which reveals that the variances of men
Fig. 7. The example of one of the features of average fluctuation $F(w)$: (a) men; (b) women.

Fig. 8. Average fluctuation $F(w)$ of men (M) and women (W).

Fig. 9. Box and whiskers graph of Hurst exponent of men’s and women’s features.
Table 4. Probability of features extraction from the DFA.

<table>
<thead>
<tr>
<th>Test</th>
<th>Width of windows (w)</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>p of F test</td>
<td>0.061545</td>
<td>0.092202</td>
<td>0.183118</td>
<td>0.335704</td>
<td></td>
</tr>
<tr>
<td>p of T test</td>
<td>0.001893</td>
<td>0.000830</td>
<td>0.000110</td>
<td>0.000003</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 10. K-means clustering plot of HR and PPI features.

and women were equal on these windows. Based on the F-test, the T-test was performed, where p-values on all windows were smaller than 0.05. Bigger windows were not considered because the value of p of T-test > 0.05. Both tests reveal that the data of men and women were significantly different on short windows and provides best results for the feature extraction. The technique indicates that the pattern behaviour and serial correlation of features data series of men and women were different in short windows.

4. Discussion

Clinically there is no significant difference in terms of blood pressure of men and women. However, after puberty, men tend to have higher blood pressure. Furthermore, after menopause women tend to have higher blood pressure than men at that age [26]. Blood pressure is influenced by various factors such as heart as the main organ of blood circulation, age and gender. [10] found considerable differences between HRV parameters of men and women.

K-means cluster analysis was used to evaluate the features of HR with PPI, PSD and DFA features using ELKI data mining software [28]. The result graphs are shown in Figure 10, 11 and 12. It shows that the features of HR with PPI, PSD and DFA were well separated into two groups of cluster memberships. The clustering result of HR with PPI is shown in Figure 10. It shows two memberships of groups for men and women and they were well separated. However, the plot also shows that HR features were identical between men and women features. The groups were separated by PPI features. Even though, the centroids were quite far from each other. Figure 11 shows the cluster memberships of PSD features. The plot shows the group membership for each band to others. It reveals that the features were grouped into two memberships. The centroids between two groups for each others were relatively far. However, there was an overlap...
of membership for Frequency Band 3 and Band 4. The final cluster centroids between the two groups were far.

Fig. 11. K-means clustering plot of PSD features.

Fig. 12. K-means clustering plot of DFA features.

The plot of K-means clustering for DFA features is shown in Figure 12. The plot shows results
identical to the PSD features plot in which the DFA features were divided into the two groups. It demonstrates that the features for each window were well grouped into two memberships. The centroids between two groups for all occasions were significantly far. Overlapping membership was not shown for all groups. The expected result was also shown in the final clustering in which the features can be distinguished into the two groups memberships.

The results demonstrate that features from all techniques could be clustered. However, PSD and DFA techniques provide more features that can provide more accurate result.

5. Conclusions

This research was conducted to determine the difference of PPG signals of men and women. The signals were analysed using PSD and DFA. As the comparison used standard features techniques, HR and PPI were calculated. The F-test and T-test were performed to distinguish the features of men and women for all techniques. There was no difference in HR features between men and women. Opposite result was shown by PPI features. The results reveal that there were significant differences features of men and women on Frequency Band 1 and Frequency Band 2 using PSD technique and short windows (4, 8, 16 and 32) using DFA. To determine the difference of the data, F-test and T-test were conducted. The F and T test demonstrated that these two groups of data, men’s and women’s features were significantly different. A clustering analysis was applied to the features from all techniques to figure out whether the features can be distinguished into groups. K-means clustering was employed to collect similar features into groups. However, PSD and DFA techniques provided more features compared to HR with PPI. The results suggest that PSD and DFA are valuable methods for extracting the features of PPG signal and that the features could be used for classification purposes.

References


