A framework for automatic text generation of trends in physiological time series data

Hadi Banaee, Mobyen Uddin Ahmed, and Amy Loutfi
Center for Applied Autonomous Sensor Systems (AASS)
Örebro University, Sweden
{hadi.banaee, mobyen.ahmed, amy.loutfi}@oru.se

***Abstract*—Health monitoring systems using wearable sensors have rapidly grown in the biomedical community. The main challenges in physiological data monitoring are to analyse large volumes of health measurements and to represent the acquired information. Natural language generation is an effective method to create summaries for both clinicians and patients as it can describe useful information extracted from sensor data in textual format. This paper presents a framework of a natural language generation system that provides a text-based representation of the extracted numeric information from physiological sensor signals. More specifically, a new partial trend detection algorithm is introduced to capture the particular changes and events of health parameters. The extracted information is then represented considering linguistic characterisation of numeric features. Experimental analysis was performed using a wearable sensor and demonstrates a possible output in natural language text.**

***Index Terms*—Health monitoring, physiological data analysis, body area networks, natural language generation, linguistic summarisation**

I. Introduction

Physiological data monitoring using body area networks (BAN) offers an opportunity to measure the quality of health conditions at more frequent intervals and outside a clinical setting [1], [2], [10], [12], [16], [17]. A main challenge for such applications is to process the large volumes of data which are produced from wearable sensors in a manner that is accessible and easily understood to both to the primary user as well as secondary users such as healthcare professionals. To facilitate proper interaction between human and the data from the body area networks, the gap between the low level sensor data and a high level symbolic representation of the data needs to be bridged. Bridging this gap involves both capturing the most significant information as well as providing informative summaries. Natural Language Generation (NLG) is one technique whereby natural language representations are generated from machine representation systems. Notable NLG systems include Reiter and Dale [14], which creates a text summary to present the nonlinguistic information in a readable text format. This NLG system has been designed to detect and summarise events in the input data, recognise the significance of information and its compatibility to the user, and then generate a natural language text which shows this knowledge in an understandable way [15]. A specific instantiation of this system for clinical data is BabyTalk project [13], which is developed to summarise clinical patient records in a neonatal intensive care unit (NICU) with considering various time scales for different end users [6]. However, most of proposed NLG systems that use physiological data have been developed using intuitive data analysis techniques which are based on expert knowledge, and predefined rules in order to detect and retrieve events in the signals. In this paper, we present a framework to generate natural language text which is representing the extracted numeric information from health parameters where unsupervised techniques are employed to detect relevant events in the data and represent them to a user. We propose a system architecture which is capable of performing textual summary for long-term physiological data. Specifically, we propose a new partial trend detection method to identify the particular changes in health signals.

The paper is outlined as follows: In Section II, we first present the data analysis module which contains several processes to produce informative features. Then we describe the NLG system to define the process of text generation from the nonlinguistic information with considering the symbolic characterisation of numeric features. Next, the progress of data collection, implementation and the sample output of the developed system is described in Section III. Finally, the last section presents conclusions and future work orientation.

II. System Architecture

This section outlines the proposed system architecture which is presented in Fig. 1. This framework analyses the input measurements and then transform the discovered numeric information to natural language text. Following parts of this section describe the modules of this architecture.

A. Input Measurements

The proposed framework has been designed to consider several continuous health parameters which are collected by wearable sensors or clinical records of physiological data. In this work we use a wearable sensing device called Bioharness3 [18] which records various vital signs on the body including heart rate, respiration rate, skin temperature, activity, and electrocardiogram (ECG). This sensor is worn on the chest and is able to locally store data or wirelessly transmitting it via Bluetooth (Fig. 2). In this architecture, the input data is one

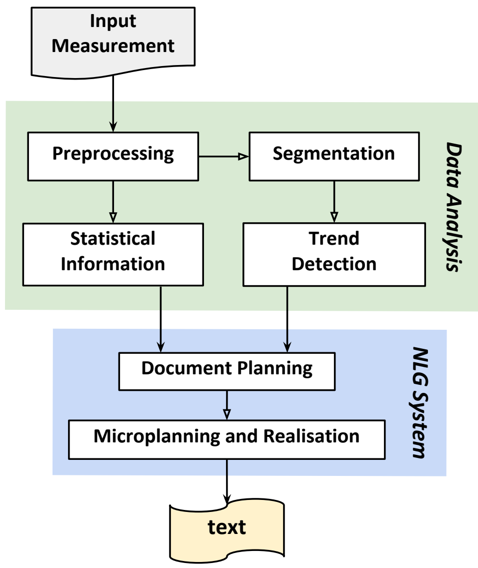

Fig. 1. System architecture of text generation from physiological time series including data analysis and NLG steps.

continuous measurement which includes a set of physiological time series for a specific period of time.

We focus primarily on two health parameters, heart rate (HR) and respiration rate (RR) which are most popular in the health monitoring domain. In this paper, each health parameter is considered independently. An example of the input measurement for the framework is shown in Fig. 3. The data in the figure has been recorded for thirteen continuous hours during the sequential activities such as exercising, walking, watching TV, and sleeping.

B. Data Analysis

The mining of physiological time series data is significant not only to model but also to detect specific health-related vital signs. One of the main challenges in healthcare area is how to analyse physiological data such that valuable information can help the end user (physician or layman). The aim of the data analysis module is to detect and represent the principal events and significant trends which are relevant for the end user. The proposed data processing method is unsupervised i.e., without expert knowledge or pre-defined rules, and can discover information which is not necessarily recognisable by an expert at first glance. As shown in Fig. 1 the data analysis

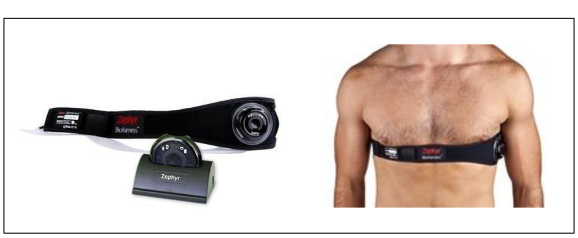

Fig. 2. The wearable sensor, Bioharness3, worn on the chest is able to locally store measured data or wirelessly transmit it via Bluetooth [18].

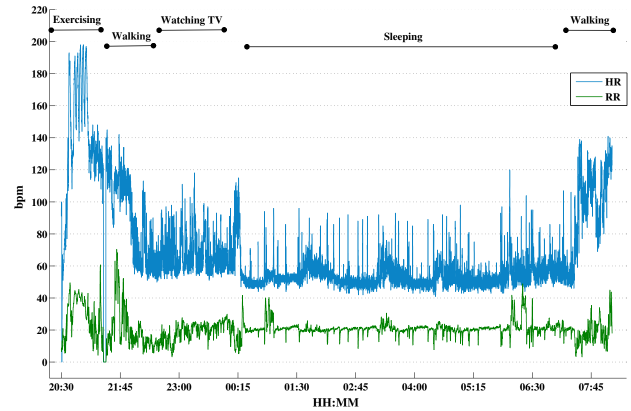

Fig. 3. An example of input measurements depicting thirteen hours of heart rate (HR, top) and respiration rate (RR, down) during sequential activities. The unit "bpm" is used for both signals (beats per minute for heart rate and breaths per minute for respiration rate).

module includes preprocessing and segmentation steps which help the system to perform statistical information and trend detection components.

1) Preprocessing: In comparison with clinical data, noise and artificats are more predominant in the signal from the wearable sensors and preprocessing of the signals is necessary. In this work, artifacts are manually removed eliminated for each health parameter. A local regression method (LOESS) has been applied to reduce the noise in the signals. This method is a nonparametric regression method which is commonly used as smoothing function [11]. Fig. 4-b shows an example of the LOESS smoothing model for heart rate and respiration rate for the raw data presented in Fig. 4-a. The bandwidth parameter in this method is adapted depending on the requirements of the output for resolution of information. The output of this step is a prepared time series data for further analysis.

2) Statistical Information: The range of information which can be discovered from physiological data is wide. However the first step to provide informative features is retrieving statistical information from input time series such as mean, mode, and variance. The average and mode of the time series values are beneficial information for the beginning of the summarised text which helps the end user to have an overview of the total measurement. For this process the system uses denoised and smoothed data to represent the underlying statistical values. To calculate the mode of the input signal, a trapezoidal fuzzy membership function is applied to find the best range of the values that includes most of the data.

3) Segmentation: After the smoothed signals are generated, a representation for each time series that captures temporal changes in the data is generated. Several methods have been introduced such as Fourier and Wavelet transforms, Symbolic Mappings, and Piecewise Linear models etc. [8] to represent the main obvious attributes of the signals. Here, piecewise linear approximation (PLA) [9] as a segmentation method is selected for this system which is able to make a significant

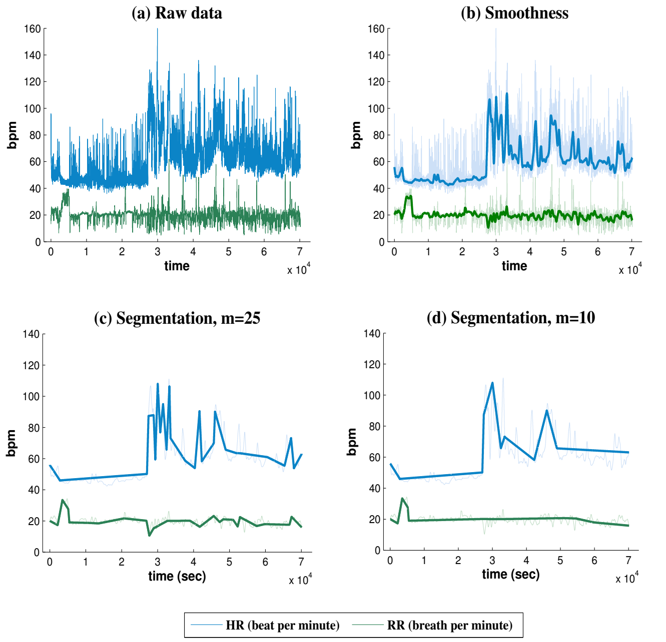

Fig. 4. An example of physiological time series data preprocessing and segmentation: (a) the raw data of heart rate (HR, top) and respiration rate (RR, down) for 22 hours, (b) an instance of smoothing raw signals using LOESS method, (c) and (d) examples of the segmentation method (PLA) of the smoothed time series in Fig. 4-b with m=25 and m=10, respectively.

representation of the time series in a simple way and efficient manner. The output of PLA method on a time series with length n is a set of linear segments with size m ($m \ll n$). The most popular approach to calculate the PLA is Bottom-Up method. This approach starts with $n/2$ segments and merges the two next segments which have minimum distance error after merging. This process repeats till some stopping criteria are satisfied. The criteria could be the threshold on maximum distance error or the number of segments.

There are several methods to find the optimal number of segments [4]. However, tuning the threshold parameter in this work is directly correlated to the resolution of information that we need to represent. The Fig. 4-c and Fig. 4-d show the examples of output of PLA method for the smoothed heart rate and respiration rate time series presented in Fig. 4-b with m=10 and m=25, respectively. Based on the request from the end user, we will be able to provide all the details of happenings during the measurements or just reporting the major trends and changes with tuning the number of segments.

4) Partial Trend Detection: Trends are an important feature to detect in physiological time series data as it can provide indications at an early stage of potential health issues and facilitate prevention. The acquired segments from the PLA approach are used to detect trends by applying a trend detection algorithm. The aim of this method is finding specific trends such as dropping, rising, and unstable changes along the measurement. Several works have previously examined partial trends on time series segmentation [3], [7]. However, the challenge here is to determine which number of segments corresponds to

Algorithm 1: Partial trend detection.

Data: Set of segments, $S = \{s_1 \dots s_m\}$
Result: Set of trends, $A = \{a_1 \dots a_l\}$, $l \leq m$
initialisation;
new trend a;
foreach s_i *belongs to* S **do**
  if $f(grad_{s_i}, len_{s_i}) > 0$ **then**
    if s_i *and* s_{i-1} *are in different gradient* **then**
      add a to A;
      new trend a;
    add s_i to a;
  else
    if $len_{s_i} < \lambda$ **then**
      if s_i *and* s_{i-1} *are in different gradient* **then**
        if s_i *and* s_{i+1} *are in same gradient* **then**
          add a to A;
          new trend a ;
          add s_i to a;
      else
        add s_i to a;
    else if $grad_{s_i} < \alpha$ **then**
      add a to A;
      new trend a ;
  go to the next s_i;
end

significant events in the data. In other words, if the number of generated segments is high and each corresponds to an event in natural language, the system will generate too many messages and overburden the user. In contrast, if the number of segments is low important information about events could be lost.

To find a proper solution to represent the trends, a partial trend is considered to be a subset of segments which have a similar *tendency* that relate to the orientation of data. So, each time series is therefore represented as a collection of *partial* trends.The preliminary step of the algorithm is to normalise the both axes of data by scaling between 0 and 1. With normalisation, the features of detected trends will be independent of the duration and range of data (long-term and short-term data).

Considering as input of the proposed trend detection method the set of segments obtained from the previous step, after normalising, the method characterises the main attributes of segments: longitude and gradient. For each segment s_i, the length of segment (len_{s_i}) and its gradient ($grad_{s_i}$), the trend detection algorithm starts with a set of segments, $S = \{s_1 \dots s_m\}$. Based on the defined parameters, the algorithm makes a decision for the segment s_i: keep it and concatenate it with the current trend, keep it as the first segment of a new trend, or ignore it. The following function has been defined to make a balance between the features of s_i:

$$f(grad_{s_i}, len_{s_i}) = (\alpha - grad_{s_i}) - 1/(\lambda - len_{s_i}) \times k$$

where the α and λ are the heuristic thresholds for gradient

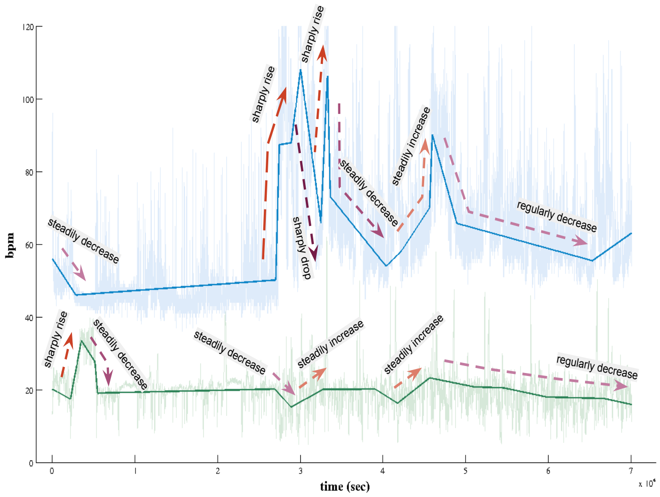

Fig. 5. The output of the partial trend detection algorithm for two segmented time series (HR, top and RR, down). The annotations of linguistic characterisation for detected partial trends are also indicated.

and length, respectively and k is a coefficient to adjust the dependency of features. In this function, if $f(grad_{s_i}, len_{s_i})$ is more than zero then s_i is kept, otherwise it will be eliminated (except some conditions related to length of s_i and the gradients of its adjacent). Algorithm1 illustrates the trend detection method with showing in which cases the algorithm makes new trend or merges the segments in the current trend.

Fig. 5 presents an output of trend detection algorithm for the segmented heart rate and respiration signals. We will describe the annotations on the detected trends in Section III.

C. NLG System

The NLG system here is inspired by the architecture proposed by Reiter and Dale [14] which includes data interpretation, document planning, microplanning and realisation to generate the text. Depending on the aim of the framework, document planning is responsible to determine which messages should appear, which messages should be combined and finally, how they should be arranged in the final text output. The role of the microplanning module is to decide which specific words should be selected to express the content, decide upon the expressions to use, and aggregate the generated messages from document planning into a linguistic structures such as paragraphs and sections. Finally, a realisation system converts the representation words and phrases in the microplanning step into an actual text considering grammatical rules. In this section we have just described the linguistic characterising of the detected trends in the data analysis process which is a part of microplanning module. For other tasks in NLG system, the framework follows the same methods in [14] and the extensions of NLG system in recent works [6].

Linguistic Characterisation of Trends: While extracting partial trends from time series data to represent them in natural language, we have a strategy to interpret the orientation of detected trends in linguistic terms. For this reason, we consider

TABLE I
THE INSTANCES OF LINGUISTIC TERMS USED FOR DESCRIBING TRENDS.

<table>
<tr><th>Duration / Range</th><th>Short</th><th>Medium</th><th>Long</th><th></th></tr>
<tr><td>Small</td><td colspan="2" rowspan="2">steadily, normally</td><td>slowly</td><td rowspan="3">Adverb</td></tr>
<tr><td>Medium</td><td>gradually</td></tr>
<tr><td>Big</td><td>suddenly</td><td>sharply</td><td>regularly</td></tr>
<tr><td>-</td><td>rise, drop</td><td colspan="2">increase, decrease, recover</td><td>Verb</td></tr>
</table>

two following features of each trend: (1) the duration of trend and (2) the range of values that trend belongs to. In order to meet the requirements of the end user and domain specificity, the system uses a fuzzy granulation. A heuristic is used to map between the mentioned features and the linguistic terms considering the following: the duration of the trend to be represented (short, medium, long), and the range of trend would be represented (small, medium, big). Note that, depending on the goal of the system, the end user's needs and type of input signal (which health parameter), the function of identifying these terms may vary.

With this categorisation, the system is able to fetch the linguistic terms to describe each trend (with specified duration and range) in natural language sentences. These sentences include particular portions such as subject, verb, adverb etc. which have to be clarified by the system. An example of defined lexicons for the trend's behaviour is illustrated in table I which includes a set of suggestions for the proper verbs and adverbs in each combination of specified duration and range. Fig. 5 shows some instances of linguistic terms for extracted trends in HR (top) and RR (down) signals.

III. RESULTS

In this section, we describe the flow of creating the proposed system which includes data collection, implementation of interface and sample output on real data sets of health parameters to see the efficiency of the generated text.

A. Data Acquisition

As previously mentioned, the input measurements have been collected by Bioharness3 wearable sensor from Zephyr Technology [18] which can measure several parameters of vital signs. For this work we used the sensor to collect long-term data from a healthy young subject. For a couple of weeks the sensor has been worn by the subject (except the times for charging device). The duration of each measurement ranges from 13 to 22 hours. The recorded data contains several health parameters such as heart rate, respiration rate, body temperature etc.. As a preliminary step to test the proposed summarising system, we considered two health parameters, heart rate (HR) and respiration rate (RR).

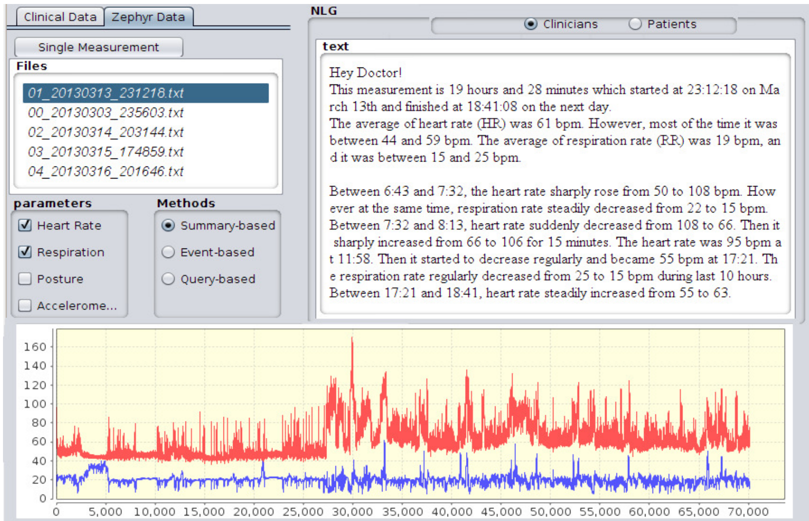

Fig. 6. A screenshot of the implemented interface for proposed framework with various options for the user to select interesting intervals, health parameters, and the audience of the system.

B. Implementation

The proposed framework has been implemented in a Java platform which is designed to support several kinds of inputs of physiological measurements. For creating a proper application for the end user, the interface consists of various options to customise the output text such as health parameters, end users of the system, and the method to define the resolution of the data. In this interface, the end user is able to select his/her interesting intervals of the input measurement as a query to the system. A screenshot of the implemented interface is shown in Fig. 6 which illustrates the final summary text for the mentioned measurement in Fig. 4. In implementing modules, for text realisation module in NLG system, we employed the SimpleNLG library [5] which is developed to generate grammatically correct English text.

C. Sample Output

To show the effectiveness of our proposed framework, several experiments are carried out. This system provides the summary text for health parameters which includes two major sections of messages: global information and trends representation. The first portion of messages in each text is global information which covers primary properties of the selected measurement such as duration, start and end time. Also the basic statistical information obtained from the data analysis module appeared in this section. An instance of the generated messages for global information of the input signals (shown in Fig. 4) is as follows.

"This measurement is 19 hours and 28 minutes which started at 23:12:18 on March 13th and finished at 18:41:08 on the next day."
"The average of heart rate (HR) was 61 bpm. However, most of the time it was between 44 and 59 bpm. The average of respiration rate (RR) was 19 bpm, and it was between 15 and 25 bpm."

Depending on the calculated average and the percentage of data in mode range, the adjectives and conjunctives for a sentence may differ. Some terms like "Most of the time" is used to show the repetition of features and some words like "However" is applied to emphasise attention to potential events.

The next portion of the final text contains the provided messages from the representation of the detected trends information. As we discussed in the previous section, after finding proper changes with a new partial trend detection algorithm, the NLG module is able to map the numeric features of the detected trends to linguistic terms. Fig. 5 presents a set of the extracted trends for both heart rate and respiration rate for the measurement shown in Fig. 4. Also the figure contains the annotations of linguistic terms for each trend. These terms are obtained from the process described in the microplanning module (table I). In the following, one instance of the provided messages for trends of input signals is presented:

"Between 6:43 and 7:32, the heart rate sharply rose from 50 to 108 bpm. However at the same time the respiration rate steadily decreased from 22 to 15 bpm."

"The heart rate was 95 bpm at 11:58. Then it started to decrease regularly and became 55 bpm at 17:21."

Another issue to mention is that depending on the end user requirements, the proposed system supported multi-resolution processing of the input signal and is able to summarise both long and short term measurements. Fig. 7 shows the output of the trend detection algorithm for two different resolutions of one measurement on heart rate. The first one is long-term data in 22 hours (Fig. 7, up) and the second one is short-term data in 4.5 hours (Fig. 7, bottom). The generated text for each resolution would be corresponded to the detected trends on that resolution. Here, the time series in the second diagram is annotated by several trends in 4.5 hours. Whereas, the same interval of data in the first diagram is just annotated as one decreasing trend.

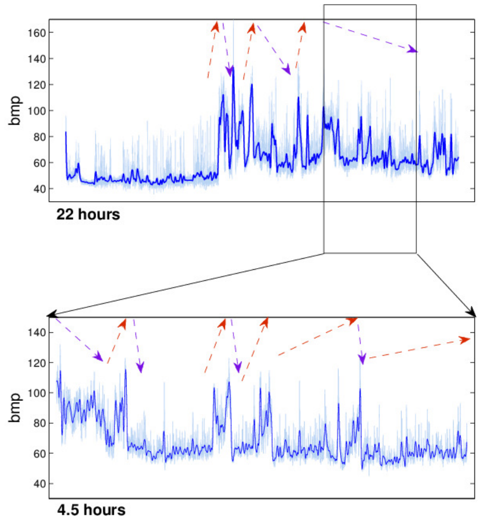

Fig. 7. An example of trend detection outputs for two different resolutions of heart rate data. Above diagram, 22 hours data with detected trends. Below diagram, 4.5 hours data captured from the last trend in the first diagram.

IV. Conclusion and Future work

This paper presented a system architecture to generate a textual summary which represents the extracted numeric information from physiological time series data. Here, the framework is capable of generating natural language text for health parameters while considering particular changes during the continuous wearable sensor-based measurements. In this paper, we proposed a partial trend detection method to identify the main events in physiological time series. After achieving nonlinguistic information, we developed the text generation approach using linguistic characterisation of the detected trends to appear in the final summary text. Furthermore, the influence of the detected trends and linguistic characterisation of numeric information is studied to assist the end user by generating readable output text. Future work in physiological data analysis may also consider the dependency of multi health parameters and recognise significant patterns and discords. The textual summary also requires further study in order to automatise the process of linguistic characterisation of numeric features and define different lexicons for various situations. Moreover, we plan to perform real-world tests of the developed framework on patients or elderly people to evaluate the generated text for both physician and layman end users of the system.

References

[1] M. U. Ahmed, H. Banaee, A. Loutfi, *Health monitoring for elderly: an application using case-based reasoning and cluster analysis*, Journal of ISRN Artificial Intelligence, In press, 2013.

[2] M. Baig, H. Gholamhosseini, *Smart health monitoring systems: an overview of design and modeling*, Journal of Medical Systems, vol. 37, no. 2, pp. 1-14, 2013.

[3] S. Boyd, *TREND: A system for generating intelligent descriptions of time-Series data*, Proceedings of Intelligent Processing Systems, 1998.

[4] T. C. Fu, *A review on time series data mining*, Engineering Applications of Artificial Intelligence, vol. 24, no. 1, pp. 164-181, 2011.

[5] A. Gatt and E. Reiter, *SimpleNLG: A realisation engine for practical applications*, Proceedings of European Natural language generation (ENLG),pp. 90-93, 2209.

[6] J. Hunter, Y. Freer, A. Gatt, E. Reiter, S. Sripada, C. Sykes, *Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse*, Journal of Artificial Intelligence in Medicine, vol. 56, no. 3, pp. 157-172, 2012.

[7] J. Kacprzyk, A. Wilbik, S. Zadrożny, *Linguistic summarization of time series using a fuzzy quantifier driven aggregation*, Journal of Fuzzy Sets and Systems, vol. 159, no. 12, pp. 1485-1499 , 2008.

[8] J. Lin, E. Keogh, S. Lonardi, B. Chiu, *A symbolic representation of time series, with implications for streaming algorithms*, Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 2-11, 2003.

[9] E. Keogh, S. Chu, D. Hart, M. Pazzani, *Segmenting time series: a survey and novel approach*, Data Mining In Time Series Databases, vol. 57, pp. 1-22, 2003.

[10] O. Lara, M. Labrador, *A survey on human activity recognition using wearable sensors*, Communications Surveys and Tutorials, IEEE, vol. PP, no. 99, pp.1-18, 2012.

[11] C. Loader, *Smoothing: local regression techniques*, Springer Handbooks of Computational Statistics, pp. 571-596, 2012.

[12] A. Pantelopoulos, N. G. Bourbakis, *A survey on wearable sensor-based systems for health monitoring and prognosis*, IEEE Transactions on Systems, Man, and Cybernetics, vol. 40, no. 1, pp. 1-12, 2010.

[13] F. Portet, E. Reiter, A. Gatt, J. Hunter, S. Sripada, Y. Freer, C. Sykes, *Automatic generation of textual summaries from neonatal intensive care data*, Journal of Artificial Intelligence, vol. 173, pp. 789-816, 2009.

[14] E. Reiter, R. Dale, *Building natural language generation systems*, Cambridge University Press, Cambridge, UK, 2000.

[15] E. Reiter, *An architecture for data-to-text systems*, Proceedings of European workshop on natural language generation, pp. 97-104, 2007.

[16] A. Wilbik, J. M. Keller, G. L. Alexander, *Linguistic summarization of sensor data for eldercare*, Proceedings of Systems, Man, and Cybernetics (SMC), pp. 2595-2599, 2011.

[17] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J. Chang, L. Hua, *Data mining in healthcare and biomedicine: a survey of the literature*, Medical Systems, vol. 36, no. 4, pp. 2431-2448 , 2012.

[18] *http://www.zephyr-technology.com*, Accessed April 10, 2013.