Blood Glucose Prediction by Breath Analysis System with Feature Selection and Model Fusion

Ke Yan and David Zhang-*IEEE Fellow*

***Abstract*— It has been shown that the concentration of acetone in breath is correlated with the subject's blood glucose level (BGL). Therefore, noninvasive BGL monitoring of diabetics can be achieved by the analysis of components in breath. In this paper, a breath analysis device with 10 gas sensors is designed to measure breath samples. Transient features are extracted from the signals of the sensors. Sequential forward selection is applied on the features to find the most informative ones. In order to reduce the interference brought by the inter-subject variance of breath acetone, global and local BGL prediction models are built and fused. The two models are based on different training strategies and have different advantages. Experiments were conducted using 203 breath samples from 36 diabetic subjects. Results show that the accuracy of the proposed feature is better than other similar features and the model fusion strategy is effective. The mean absolute error and mean relative absolute error of the system are 2.07 mmol/L and 20.69%, respectively.**

I. INTRODUCTION

The frequent monitoring of blood glucose levels (BGLs) is important for diabetics. However, typical BGL measurement devices use automatic lancets to prick the fingertip for blood samples, which is invasive and painful. Noninvasive BGL monitoring approaches including reverse iontophoresis, bioimpedance spectroscopy, near infrared spectrophotometry, and so on [1] have been studied. They provide painless and convenient measurements, but still suffer disadvantages such as being sensitive to environmental variations, subjects' skin conditions and movement [1].

Diabetics are patients whose body cells cannot absorb the glucose in blood properly. When their livers break down fat for energy, the ketone bodies in their blood will increase [2]. Acetone is one of the three kinds of ketone bodies. Since it is volatile, it can be detected in the exhaled breath. Lots of researchers have studied the relationship between breath acetone and BGL. They showed that the concentration of acetone in breath is correlated with BGLs [3][4]. As a result, the analysis of components in breath can be used as an assistive noninvasive approach for BGL monitoring on both hyper- and hypoglycemic patients [3][4].

The measurement of components in breath is usually performed by gas chromatography combined with mass spectrometry (GC/MS). Recently, chemical sensor systems, also known as electronic noses (e-noses), are attracting more and more attention. Compared to GC/MS, they are cheaper, faster, more portable and easier to operate. With the development of gas sensor technology, their precision is becoming more and more satisfactory. They have been successfully applied in medicine for diagnosis of diabetes, renal disease, airway inflammation [5], and so on.

In the last few years, some researchers have applied e-noses on BGL monitoring. Guo et al. proposed a breath analysis device with 12 metal oxide semiconductor (MOS) sensors. Principal components were extracted from the magnitude of the sensors' signals. Support vector ordinal regression was used to classify 192 diabetics into 4 groups according to their BGLs [6]. An accuracy of 68% was achieved. The authors of [7] built up an e-nose to predict the BGL of 30 diabetics. The frequency differences and the first values of 6 sensors were adopted as inputs to a neural network (NN). The BGLs were obtained from the outputs of the NN, which had a mean relative absolute error of 25.24%.

There are two problems in the previous studies. First, the features in both studies were not optimized. Features are computed from the sensors' signals to extract information. Both studies explored only the magnitude of the sensors' signals. In fact, transient features such as the derivative, integral and time feature of the signals may contain more meaningful information in gas sensors [8][9]. For example, the recovery time of a sensor is related to the type of the analyzed gas [9]. The dimension of transient features can be very high. Feature selection algorithms are often used to remove irrelevant and redundant features while keeping only the useful ones [9][10]. By properly selecting a subset of transient features, the computational complexity of the algorithm can be reduced while the accuracy can be improved.

The other problem is caused by the inter-subject variance of breath acetone. As shown in [4], although breath acetone is correlated with BGL for each subject, its baseline values vary among subjects. The author of [2] concluded that calibration of acetone with BGL for each individual is required. However, global BGL prediction models for all subjects were built in previous studies [6][7]. This model development strategy should be improved for better prediction accuracy.

In this paper, a breath analysis device (e-nose) with 10 gas sensors is designed to measure the breath of diabetics. Transient features are extracted from the sensors' signals.

This work was partially supported by the Natural Science Foundation of China (NSFC) (No. 61332011, 61020106004, 61272292, 61271344), the GRF fund from the HKSAR Government, the central fund from Hong Kong Polytechnic University, Shenzhen Fundamental Research fund (JCYJ20130401152508661), and Key Laboratory of Network Oriented Intelligent Computation, Shenzhen, China.

K. Yan is with the Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China (e-mail: yank10@mails.tsinghua.edu.cn).

D. Zhang is with the Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China, and also with the Department of Computing, Biometrics Research Centre, The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: csdzhang@comp.polyu.edu.hk).

Sequential forward selection is adopted to select a subset of features which have high precision on BGL prediction. Experiment results show that the selected subset is more effective. Based on these features, two kinds of prediction models are proposed. The global model is trained using breath samples from all subjects with an additional feature to indicate the subject's identity, while the local model only uses the samples from one subject as training samples. By fusing the two models in score level, the prediction accuracy is further enhanced.

II. METHODS

A. Breath Analysis System

The framework of the proposed system is displayed in Fig. 1. When collecting a breath sample, the subject is asked to exhale into a Tedlar gas bag. Then the breath gas is drawn from the gas bag into a gas room which contains an array of sensors. The signals of the sensors are captured by a signal processing circuit. They are filtered and amplified, then transmitted into a computer. The stored digitized breath sample includes response curves from the 10 sensors.

The sensor array consists of a carbon dioxide sensor and 9 MOS sensors sensitive to volatile organic compounds (VOCs) [11]. The VOC sensors were specially selected commercial sensors for the analysis of diabetics' breath. They have different sensitive spectrums and measurement ranges so as to provide complementary information. Some of them are particularly sensitive to acetone. Three of the sensors are under temperature modulation, a technique that can enhance the discriminative power of MOS sensors [12]. A staircase modulation voltage [12] is applied to the three sensors.

After a digitized breath sample is acquired, it undergoes a series of data analysis algorithms before the BGL estimation is got. In the preprocessing step, the baseline value in each sensor's response is subtracted [5]. Transient features are then extracted and selected. The feature extraction and selection methods will be introduced in Section II.B and II.C, respectively. Finally, a global regression model and a local regression model are developed based on different training samples. The two models use different feature subsets which are selected separately. The outputs of the two models are fused to get the BGL prediction. Section II.D will describe the model development and fusion strategies.

B. Transient Feature Extraction

Fig. 2 illustrates a preprocessed sample from a diabetic subject. S1-S7 are sensors without temperature modulation (TM). Their response curves have three main stages. In the injection stage, the sensors are exposed to breath and their responses start to rise. In the reaction stage, the sensors are in full contact with breath and the responses reach their peak levels. In the purge stage, the gas room is purged with clean air and the responses gradually return to baseline. S8-S10 are sensors with TM. Their response curves oscillate in a staircase manner along with the modulated temperature.

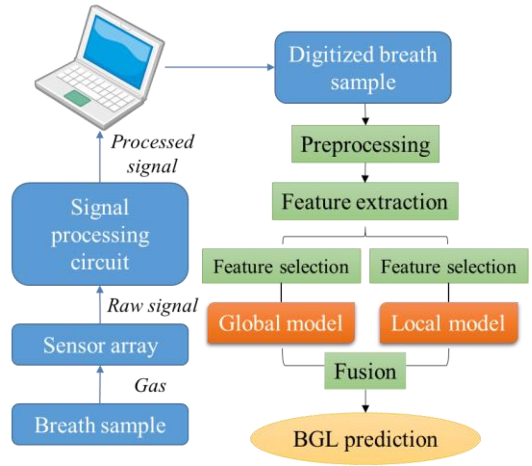

Figure 1. The framework of the proposed breath analysis system.

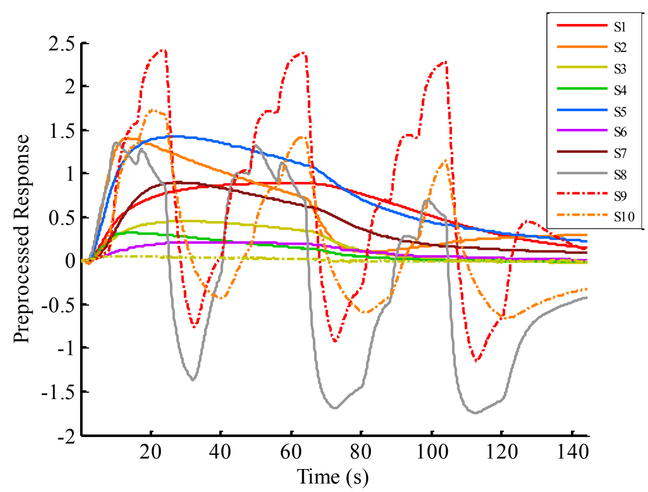

Figure 2. A preprocessed sample from a diabetic subject.

For sensors both with and without TM, rich information about the type and concentration of the analyzed gas is in their transient features. Six kinds of transient features are explored in our experiments: magnitude, difference, derivative, 2nd order derivative, integral and time features. For each sensor without TM, 69 transient features are extracted from its response curve; for each sensor with TM, 208 features are extracted since the curve contains more shape information. Some examples of the transient features are:

- Down-sampled magnitude;
- Maximum and minimum derivative;
- The time when the magnitude reaches 70%, 50% and 30% of the maximum magnitude in the purge stage.

C. Feature Selection

After combining transient features from the 10 sensors into a feature vector, the dimension of the vector reaches 1107. This feature size is too high for regression algorithms to learn accurate models, especially when the training samples are not sufficient. We use sequential forward selection (SFS) to remove irrelevant or redundant features and keep only the useful ones. SFS is an algorithm with moderate computational complexity and relatively good accuracy [9][13]. It is a greedy method. In the first step, one feature is selected which alone provides the best accuracy in the prediction task. In the next step, a new feature is selected if the combination of the previous selected subset and the new feature achieves the best accuracy. This step is repeated until all the features have been selected or the accuracy stops improving. This method treats the prediction algorithm as a black box, so the user can choose the most suitable algorithm. It is more favorable when the optimal feature subset is small.

D. Prediction Model Development and Fusion

Two different regression models are developed for BGL prediction. In the global model, samples from all subjects are included in the training set. This model can make use of the most training samples. However, the inter-subject variance of acetone concentration is not considered, so its accuracy will somehow be influence. To improve this model, we further add a categorical feature in each feature vector to indicate the subject's identity. Concretely, for each test sample in the database, a regression model is trained using all the other samples. For each training sample, the additional categorical feature will be 1 if the training sample is from the same subject with the test sample, or be 0 otherwise. The test sample will also have the additional feature with the value 1. The advantage of this method is that all the training samples can contribute to the regression model while the training samples from the same subject with the test sample will be emphasized. Thus, the interference brought by the inter-subject variance can be reduced.

On the other hand, the local model is trained using only the samples from the same subject with the test sample. This model is more subject-specific. The regression algorithm can learn the relationship between breath acetone and BGL of the subject more precisely. But when the samples collected for the subject is insufficient, the regression algorithm cannot get enough training samples, so the learned model will also be inaccurate. In that case, the accuracy of the global model may be better. Therefore, we use a simple fusion strategy to combine the two models. Suppose the output of the global and local model are y_g and y_l, respectively. Then the final BGL prediction y can be expressed by (1), where w_g is the weight for the global model. It can be decided by experiments. The support vector regression (SVR) [14] algorithm is used as the regression algorithm to predict BGL in both models, since it is not prone to over-fitting.

$$y = w_g y_g + (1 - w_g) y_l \qquad (1)$$

III. Results and Discussion

In this study, 203 breath samples were collected from 36 volunteer patients in Guangdong Provincial Hospital of Traditional Chinese Medicine (Guangzhou, China). They are all Type 2 diabetes patients (17 males, 19 females; age 57.7±9.1). For each subject, several breath samples were collected at 2h after meal in different days together with the simultaneous BGLs. The number of samples per subject ranges from 3 to 10. The relationship between breath acetone and BGL of 3 subjects is drawn in Fig. 3. It can be observed that the relationship between breath acetone and BGL varies among subjects. Besides, there is also intra-subject variance, which may be due to individual influential factors such as diet [15] and insulin [2]. The subjects were all receiving 1-3 kinds of treatments including oral medication, insulin injection, and insulin pump. None of the treatments is observed to correlate with both the BGL and the breath samples.

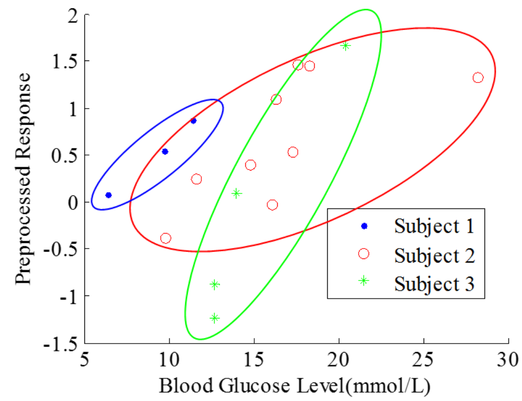

Figure 3. The relationship between breath acetone and BGL of 3 subjects. The y-axis is the maximum value of the preprocessed response of a sensor (S3) which has high sensitivity to acetone.

The samples were randomly divided into 3 subsets. SFS was applied on two of them (selection set) to select features. In the procedure of feature selection, the leave-one-out cross-validation protocol was used to judge the performance of each feature combination. Then the selected features were adopted to build models for each sample in the remaining subset (test set) to predict the BGL of the sample. The mean average error (MAE) of all samples was computed after all 3 subsets had been the test set once.

When deciding the number of selected features in the models, we did not use the number when the MAE of the selection set is minimum. To prevent over-fitting on the selection set, only the first 3 features selected by SFS were adopted to build global and local prediction models. Then the outputs of the two models are fused using (1). The best accuracy comes when $w_g = 0.6$, indicating that the two models make approximately equal contributions in the fused model. The MAE of our method is shown in Table I. MAE of several other features in global, local, and fused models are also listed in Table I for comparison.

TABLE I. MAE Comparison for Different Methods

Method	Global	Local	Fused
Maximum magnitude	2.578	4.255	2.533
Magnitude + PCA	2.324	3.070	2.275
All transient features	4.239	2.349	2.346
Transient features + PCA	2.321	2.352	2.195
Transient features + SFS	**2.262**	**2.332**	**2.072**

According to Table I, the selected subset of only 3 features outperforms the whole set of transient features. This is because the whole feature set has high dimensionality and many redundant and noise features, which interfered the learning of accurate regression models. Another commonly used dimension reduction algorithm, PCA, was also tested in our experiments. When PCA was applied to transient features, the MAE was reduced, but still higher than transient features + SFS. This is because the principal components can preserve the largest variance in original features, but they are not necessarily good for regression. Using the same model development and fusion strategy, we also tested two other feature extraction algorithms, i.e. the maximum magnitude (MM) feature similar to [7] and the magnitude + PCA feature as in [6]. The accuracy of these features was not as high as ours. These results prove that among the transient features, some features are more effective than MM and PCA. Meanwhile, SFS has the ability to select them.

The fusion of global and local models further improved the accuracy of the proposed method. This is probably because that the two models are somehow complementary. In fact, the fused model improved the performance of all the tested features. The final MAE, mean relative absolute error [7] and correlation coefficient of the proposed method are 2.072, 20.69%, and 0.6982, respectively. The values of true and predicted BGLs are plotted in Fig. 4.

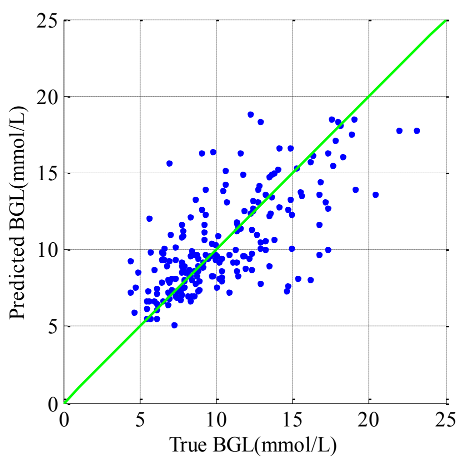

Figure 4. Scatter plot of true and predicted BGLs.

IV. Conclusion

In this paper, a breath analysis device and several data analysis methods are proposed for blood glucose level (BGL) prediction of diabetics. To find informative features from the gas sensors' signals, sequential forward selection is applied on transient features. In order to reduce the interference brought by the inter-subject variance of breath acetone, we propose to fuse global and local BGL prediction models. Experiment results have confirmed the effectiveness of these strategies. This paper illustrates the possibility of improving the accuracy of breath analysis systems according to the results in sensor and medical researches. This system has the potential to become a noninvasive and convenient tool to assist BGL monitoring. Its current accuracy is not quite high to fit the requirement of clinical use, though. Our future works will focus on methods to further enhance its performance. Intra-subject variance factors such as diet [15] and insulin [2] should be considered. More samples should be collected from each subject to develop better local models. The sequential forward selection algorithm may be trapped by local minima, so advanced feature selection algorithms should be investigated.

Acknowledgment

The authors would like to thank the volunteer patients in Guangdong Provincial Hospital of Traditional Chinese Medicine for their help on our research.

References

[1] S. K. Vashist, "Non-invasive glucose monitoring technology in diabetes management: A review," *Anal. Chim. Acta,* vol. 750, pp. 16-27, Oct. 2012.

[2] C. Turner, "Potential of breath and skin analysis for monitoring blood glucose concentration in diabetes," *Expert Rev. Mol. Diagn.,* vol. 11, no. 5, pp. 497-503, June 2011.

[3] C. Wang, A. Mbi, and M. Shepherd, "A study on breath acetone in diabetic patients using a cavity ringdown breath analyzer: Exploring correlations of breath acetone with blood glucose and glycohemoglobin a1c," *IEEE Sens. J.,* vol. 10, no. 1, pp. 54-63, Jan. 2010.

[4] C. Turner, C. Walton, S. Hoashi, and M. Evans, "Breath acetone concentration decreases with blood glucose concentration in type I diabetes mellitus patients during hypoglycaemic clamps," *J. Breath Res.,* vol. 3, no. 4, pp. 046004, Dec. 2009.

[5] D. Guo, D. Zhang, N. Li, L. Zhang, and J. Yang, "A novel breath analysis system based on electronic olfaction," *IEEE Trans. Biomed. Eng.,* vol. 57, no. 11, pp. 2753-2763, Nov. 2010.

[6] D. Guo, D. Zhang, L. Zhang, and G. Lu, "Non-invasive blood glucose monitoring for diabetics by means of breath signal analysis," *Sens. Actuators: B. Chem.,* vol. 173, no. 6, pp. 106-113, Oct. 2012.

[7] H. M. Saraoglu, A. O. Selvi, M. A. Ebeoglu, and C. Tasaltin, "Electronic Nose System Based on Quartz Crystal Microbalance Sensor for Blood Glucose and HbA1c Levels From Exhaled Breath Odor," *IEEE Sens. J.,* vol. 13, no. 11, pp. 4229 - 4235, Nov. 2013.

[8] C. Distante, M. Leo, P. Siciliano, and K. C. Persaud, "On the study of feature extraction methods for an electronic nose," *Sens. Actuators: B. Chem.*, vol. 87, pp. 274-288, Dec. 2002.

[9] N. Paulsson, E. Larsson, and F. Winquist, "Extraction and selection of parameters for evaluation of breath alcohol measurement with an electronic nose," *Sens. Actuators: A. Phys.*, vol. 84, pp. 187-197, Sep. 2000.

[10] J. H. Cho and P. U. Kurup, "Decision tree approach for classification and dimensionality reduction of electronic nose data," *Sens. Actuators: B. Chem.*, vol. 160, pp. 542-548, Dec. 2011.

[11] K. Yan, D. Zhang, D. Wu, H. Wei, and G. Lu, "Design of a Breath Analysis System for Diabetes Screening and Blood Glucose Level Prediction," *IEEE Trans. Biomed. Eng.*, to be published.

[12] A. Amini and V. Ghafarinia, "Utilizing the response patterns of a temperature modulated chemoresistive gas sensor for gas diagnosis," in *IOP Conf. Series: Mater. Sci. Eng.*, 2011, p. 012048.

[13] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157-1182, Mar. 2003.

[14] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Stat. Comput.,* vol. 14, no. 3, pp. 199-222, 2004.

[15] P. Španěl, K. Dryahina, A. Rejšková, T. W. Chippendale, and D. Smith, "Breath acetone concentration; biological variability and the influence of diet," *Physiol. Meas.,* vol. 32, no. 8, pp. N23, Aug. 2011.