Journal of
Biomedical Optics

BiomedicalOptics.SPIEDigitalLibrary.org

Screening of patients with bronchopulmonary diseases using methods of infrared laser photoacoustic spectroscopy and principal component analysis

Yury V. Kistenev
Alexander I. Karapuzikov
Nadezhda Yu. Kostyukova
Marina K. Starikova
Andrey A. Boyko
Ekaterina B. Bukreeva
Anna A. Bulanova
Dmitry B. Kolker
Dmitry A. Kuzmin
Konstantin G. Zenov
Alexey A. Karapuzikov

SPIE.

Screening of patients with bronchopulmonary diseases using methods of infrared laser photoacoustic spectroscopy and principal component analysis

Yury V. Kistenev,[a,b,*] **Alexander I. Karapuzikov,**[c] **Nadezhda Yu. Kostyukova,**[c] **Marina K. Starikova,**[c] **Andrey A. Boyko,**[c] **Ekaterina B. Bukreeva,**[a] **Anna A. Bulanova,**[a] **Dmitry B. Kolker,**[c,d] **Dmitry A. Kuzmin,**[a] **Konstantin G. Zenov,**[c] **and Alexey A. Karapuzikov**[c]
[a]Siberian State Medical University, Medico-Biological Faculty, Physics Department, Moscowski Trakt 2, Tomsk 634050, Russia
[b]Tomsk State University, University Administration, 36 Lenina Avenue, Tomsk 634050, Russia
[c]Special Technologies Ltd., Zeljonaja Gorka Street 1/3, Novosibirsk 630060, Russia
[d]Novosibirsk State University, 2 Pirogova Street, Novosibirsk 630090, Russia

Abstract. A human exhaled air analysis by means of infrared (IR) laser photoacoustic spectroscopy is presented. Eleven healthy nonsmoking volunteers (control group) and seven patients with chronic obstructive pulmonary disease (COPD, target group) were involved in the study. The principal component analysis method was used to select the most informative ranges of the absorption spectra of patients' exhaled air in terms of the separation of the studied groups. It is shown that the data of the profiles of exhaled air absorption spectrum in the informative ranges allow identifying COPD patients in comparison to the control group. [DOI: 10.1117/1.JBO.20.6.065001]

Keywords: exhaled air; chronic obstructive pulmonary disease; noninvasive express diagnostics; volatile organic compounds; laser photoacoustic spectroscopy; principal component analysis.

Paper 140861SSPRR received Dec. 25, 2014; accepted for publication May 12, 2015; published online Jun. 3, 2015.

1 Introduction

Noninvasive diagnostics are one of the most important directions for the development of modern medicine. Recently, the interest has been focused on patients' exhaled air research as a noninvasive diagnostic method for bronchopulmonary, cardiovascular, gastrointestinal, and other diseases.[1,2] The basis of similar methods is related to the variation of concentrations of volatile organic compounds (VOCs) in exhaled air according to specific diseases.[3] For example, it is ascertained that bronchial asthma exacerbation is characterized by ammonia (NH_3) concentration increasing in exhaled air by 250 to 300 times.[3,4] A high level of propane (C_3H_8) in the exhaled air was identified for different clinical forms of pulmonary tuberculosis.[5,6] The issue is significant because this approach, due to the sparse sampling, precludes pain and physical and emotional discomfort of the patient, the possibility of transmission of the blood-borne infections, and provides safety for the diagnostic studies. On one hand, noninvasive diagnostic methods can be used on an outpatient basis—this provides their widespread application; on the other hand, for patients at resuscitation departments, as the severity of the patient's state is not a contraindication for their application.

In this paper, we discuss the abilities of the methods of infrared (IR) laser spectroscopy and the principal component analysis (PCA) technique for a totally noninvasive express diagnostic of chronic obstructive pulmonary disease (COPD) on the basis of absorption spectra analysis of the patient's exhaled air.

2 Techniques and Methods

The method of laser photoacoustic spectroscopy (LPAS) is convenient for measuring the concentration of VOCs in the exhaled air because of the simplicity of its practical implementation, safety, cost-effectiveness, and extremely high sensitivity (minimal measured concentration for some chemicals at atmospheric conditions is about 1 ppb). LPAS has advantages in the detection of the gases which have overlapping of the absorption lines with the lasing lines (e.g., CO or CO_2-laser). LPAS is based on the generation of acoustic waves in a gas excited by the modulated laser beam at the wavelength corresponding to the absorption line of the VOCs in gaseous samples and on the measurement of the parameters of these acoustic waves using sensitive microphones.

2.1 *LPAS Combined with CO_2-Laser*

CO_2-lasers are one of the most suitable radiation sources for photoacoustic detection because of their narrow lasing line and commercial availability. Waveguide CO_2 laser exited by radio frequency generator (144 MHz, 1 kW) with a wavelength tuning from 9.2 to 10.8 μm was developed by Special Technologies Ltd. Selection of the necessary lasing lines was provided by diffraction grating. We realized gas analyzers based on the LPAS method related with the mentioned above CO_2-laser with intracavity (ILPA) and extracavity (LGA-2) photoacoustic detector location in cooperation with the Institute of Laser

*Address all correspondence to: Yury V. Kistenev, E-mail: yuk@iao.ru

1083-3668/2015/$25.00 © 2015 SPIE

Physics SB RAS and Institute of Atmospheric Optics SB RAS. Both devices have high sensitivity (over 1 ppb level) and spectral resolutions (over 0.003 cm^{-1}) at pulsed mode. Main technical parameters of the developed gas analyzers are shown in the Table 1.

2.2 LPAS Combined with Optical Parametric Oscillator

Frequency conversion using optical parametric oscillator (OPO) is one of the effective ways to generate widely tunable coherent light in a spectral range from visible to mid-IR ranges. These laser sources play a particularly important role in the IR range, where VOCs have their fundamental absorption lines. This is because of the OPOs' ability to provide continuous wavelength tuning over a wide spectral range[7] that the PAS method combined with OPO allows concentration determination of a number of different gases.

We developed the gas analyzer LaserBreeze based on the LPAS method and OPO with a tuning range from 2.5 to 10.7 μm. The main technical characteristics of the LaserBreeze gas analyzer are shown in Table 2. We used two types of nonlinear elements in the optical scheme: periodically poled lithium niobate structure (PPLN) and mercury thiogallate crystal $HgGa_2S_4$ (HGS). Special cavities were designed for each element and an Nd:YLF laser (10 ns, 0.5 to 1.5 kHz, 1.5 mJ) was used as a pump source. The linewidth of the developed OPOs was 3 to 4 cm^{-1}. The average power of OPO based on PPLN structure was 20 mW (1700 Hz). The average power of OPO based on HGS crystal was 9 mW (900 Hz). The double-channel resonant photoacoustic cell was used for recording the absorption spectra of gaseous samples. The LaserBreeze gas analyzer is described in detail in Ref. 8.

The main VOCs that can be detected by gas analyzers based on the LPAS method with a CO_2-laser and OPO are shown in Table 3. The sensitivity for measured gases was not lower than 1 ppb.

Given in Table 3 wavelengths were chosen so that they are out of the absorption lines of other gases in the mixture or, if it is impossible, absorption of other gases is minimized. Moreover, for some gases, it is reasonable to carry out concentration measurements in two so-called spectral measurement channels:[9] the wavelength of one spectral channel is close to or coincides with the center of one of the absorption lines of the gas (usually the

Table 1 Main technical parameters of the developed gas analyzers.

Parameters	ILPA	LGA-2
Spectral range (μm)	9.2 to 10.8	9.2 to 10.8
Lasing lines numbers	60	50
Pulse repetition rate (Hz)	1170 ± 50	170 ± 0.5
Average power (W)	0.5	1

Table 2 Main technical characteristics of the LaserBreeze gas analyzer.

Parameter	Value
Concentration sensitivity	No worse than 1×10^{-3} ppm
Number of detected molecular biomarkers	No less than 20
Relative error in determining volatile organic compounds (VOCs) concentration	No more than 10% to 30%
Reliability and selectivity of VOCs identification	No less than 95%
Scanning range of optical parametric oscillator (OPO) radiation	2.5 to 10.7 μm
Sample volume	No more than 50 cm^3
Detection time for one VOCs in a sample	No more than 3 s
Detection time for 10 VOCs in a sample	No more than 2 min

Table 3 Main VOCs.

	Possibility of detecting different VOCs by gas analyzers (the absorption bands)	
VOCs	Based on OPO (2.5 to 10.7 μm)	Based on CO_2 (9.2 to 10.8 μm)
Acetone (C_3H_6O)	7.35 μm	—
Acetylene (C_2H_2)	3.05 μm	—
Ammonia (NH_3)	10.35 μm	10.35 μm; 10.73 μm
Butane (C_4H_{10})	3.387 μm	10.45 μm
Carbon dioxide (CO_2)	4.24 μm	10.6 μm
Carbon dioxide (13 isotope) ($^{13}CO_2$)	4.408 μm	—
Carbon monoxide (CO)	4.62 μm	—
Ethane (C_2H_6)	3.348 μm	—
Ethanol (C_2H_5OH)	9.38 μm	9.38 μm
Ethyl acetate ($C_4H_8O_2$)	8.03 μm	9.47 μm
Ethylene (C_2H_4)	10.53 μm	10.53 μm
Methane (CH_4)	7.7 μm	—
Nitrogen dioxide (NO_2)	6.25 μm	—
Nitrogen oxide (NO)	5.25 μm	—
Nitrous oxide (N_2O)	3.89 μm	—
Pentane (C_5H_{12})	3.372 μm	—
Propane (C_3H_8)	3.375 μm	10.8 μm
Sulfur dioxide (SO_2)	7.28 μm	—

most intense). The spectral channel is located on the edge of the absorption line. Due to the overlapping spectra of the individual gas components, the task of selecting the spectral measurement channels becomes complex and sometimes ambiguous.[9] In this case, the use of special computational algorithms capable of selecting a set of spectral measurement channels where the errors of recovery of the gas concentrations would be minimal or close to the minimum. We used this approach for selecting the wavelengths at which the concentrations were measured.

2.3 *Principal Component Analysis and Data Preprocessing*

The measuring of VOCs' concentrations in exhaled air is a promising tool for diagnostics in the future, but it should be pointed out that a significant part of the VOCs is not highly specific. For example, asthma causes the essential growth of exhaled NO and moderate growth of CO, COPD causes a small NO growth and essential growth of CO.[3] Additionally, taking into account the individual variability in metabolism, it is obvious that for the diagnostics it is more expedient to use the "profile" of the set of VOCs or to directly profile the absorption spectrum of a breath sample as a "fingerprint" of the patient's medical condition without component analysis of the sample. In this situation, various methods of data mining promise to be effective to analyze data.[10,11] One of them is the PCA.[12,13]

The basic idea of PCA is to find the minimum number of new features that are enough for the recovery of the basic features by linear transformation, possibly with insignificant errors. PCA projects correlated variables into a lower number of uncorrelated variables called principal components (PCs). A specific feature of PCA is that the hidden connections and patterns that are typical for the investigated data set can be revealed.

The mathematical background of PCA consists of the decomposition of initial experimental data two-dimensional (2-D) matrix $X(I \times J)$ into the form of a matrix product[11]

$$X = T \cdot P^t + E = \sum_{a=1}^{A} t_a \cdot p_a^t + E, \qquad (1)$$

where I is the quantity of samples of experimental data, J is the quantity of the features of investigated objects, $T(I \times A)$ is the score's matrix, $P(A \times J)$ is the loading's matrix, E is the residual's matrix, and A is the quantity of PCs. In our case, these features of the state under investigation are the set of absorption coefficients of the exhaled air sample in the laser source frequency detuning branch of the used gas analyzer.

The loading's matrix contains weight coefficients which characterize the contribution of features to a specific PC. The score's matrix contains coordinates of the samples in the space of PCs.

PCA is useful if $A \ll J$. In this case, the method allows, first of all, to separate the most informative features of the state, or in other words, to reduce the dimensions of the feature space and to decrease noise, and second, to estimate the relative position of the studied objects in the reduced space of PCs.

3 Results and Discussion

The experimental research was carried out according to the principles of Good Clinical Practices. The protocol of the research was approved by the Ethic Committee of the Siberian State Medical University (Tomsk, Russia), Ref. Number 2882 on 24 November 2011. All participants were informed about details of the research and signed "Informed agreement" for the actions carried out. The study involved 11 healthy nonsmoking volunteers (control group) and seven patients with COPD (target group). The COPD patients were males with verified diagnoses who passed treatment at the Pulmonological Division of the Regional State Autonomous Institution of Public Health "Municipal Clinical Hospital No. 3" (Tomsk, Russia). The average age of this group was 59.6 years. We did not included COPD patients with an unverified diagnosis, the presence of pneumonia, asthma, and other respiratory pathologies. The control group consisted of conventionally healthy nonsmoking male volunteers. Inclusion criteria were the absence of acute illness within two weeks prior to sample collection, without chronic pathologies of bronchopulmonary, cardiovascular, digestive, urinary and reproductive systems, and the absence of the factor "smoking" in anamnesis vitae. The average age in this group was 21.1 years.

The procedure of exhaled air sampling was as follows. All samples were taken before eating or 2 h thereafter. The air was collected in standard test tubes. Prior to sampling, participants rinsed their mouth with running water. The study does not imply special cleaning of the oral cavity. Then the participant did some calm breaths through a sterile plastic tube into the test tube, which was then sealed with a sterile cotton wad.

All exhaled air samples were analyzed using the LGA-2 and the LaserBreeze LPAS gas analyzers. Five scans of the absorption spectrum of each sample were recorded and averaged to reduce random errors.

Most informative subranges of the measured profiles of the absorption spectra were determined by PCA. The criterion was the best spatial separation of the target group from the control group in the space of the PC. The results below are focused on the most informative subranges.

The number of PC (in other words, the dimensions of the above-mentioned space) is usually chosen to describe at least 70% of the variation of initial data [it is the so-called explained variance (EV)].[11] Here, initial data involved in the absorption

Table 4 The dependence of the explained variance (EV) on the quantity of the used principal components (PCs).

Spectral subrange (μm)	9.2 to 9.8	2.59 to 2.817	3.272 to 3.498	3.499 to 3.725
Type of gas analyzer	LGA-2	LaserBreeze	LaserBreeze	LaserBreeze
EV for second PCs	84.8%	98%	70%	86.6%
EV for third PCs	95.8%	98.9%	77.6%	89.1%

Table 5 The quantity of informative absorption coefficients.

Spectral subrange (μm)	9.2 to 9.8	2.59 to 2.817	3.272 to 3.498	3.499 to 3.725
Initial quantity	30	215	215	215
The first PC	24	159	163	180
The second PCs	2	5	19	7

spectra of all exhaled air samples. The value of EV in the used spectral subranges that is dependent on the quantity of the PC is presented in Table 4. According to the data from Table 4, the 2-D space of PC is enough to analyze profiles of the absorption spectra of the exhaled air samples.

To select the most informative set of absorption coefficients, we apply the method that is similar to the well-known "method of broken sticks" to the loading matrix.[14] The results are shown in Table 5. Here, the initial quantity is the quantity of absorption coefficients which were contained in the definite spectral subrange before PCA application.

According to the PCA, every sample is represented by the point in the space of PC. We used the freeware "ViDaExpert"[15] to estimate the spatial distribution of the exhaled air samples in the space of the PC.

The results of point estimates of the exhaled air absorption spectra profiles of COPD patients and healthy volunteers in the 9.2 to 9.8 μm subrange are shown in Fig. 1. The distance between the point estimates on the plane of the PC characterized the difference in the absorption spectra profiles of participants. This is caused by variations in metabolism and, hence, is the difference in the VOCs profile of the samples.

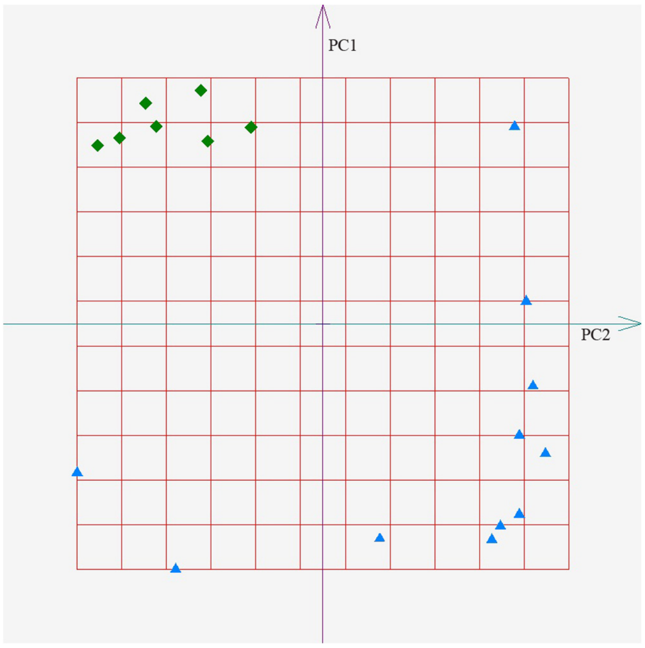

Fig. 1 Spatial distribution of the exhaled air samples from chronic obstructive pulmonary disease (COPD) patients (the diamond icons) and healthy volunteers (the triangle icons). The feature set includes absorption coefficients of the sample in the range of 9.2 to 9.8 μm. The axes correspond to the first (PC1) and the second (PC2) principal components.

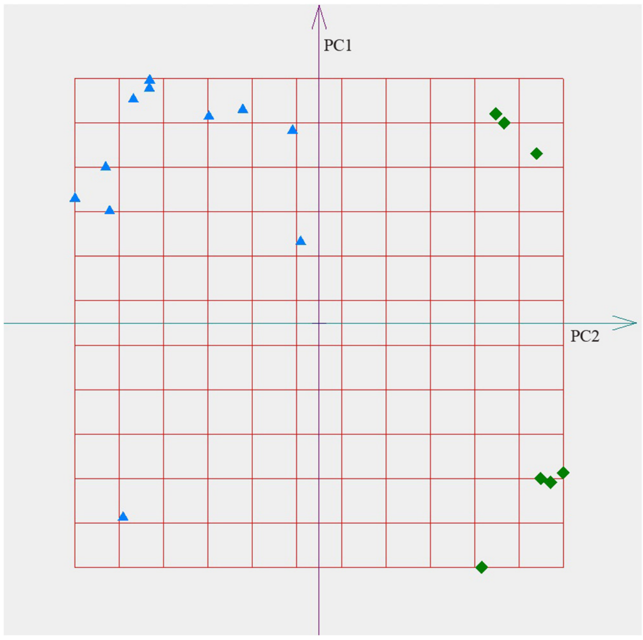

Fig. 2 Spatial distribution of the exhaled air samples from the COPD patients (the diamond icons) and healthy volunteers (the triangle icons). The feature set includes absorption coefficients of the sample in the range of 2.59 to 2.817 μm. The axes correspond to the first (PC1) and the second (PC2) principal components.

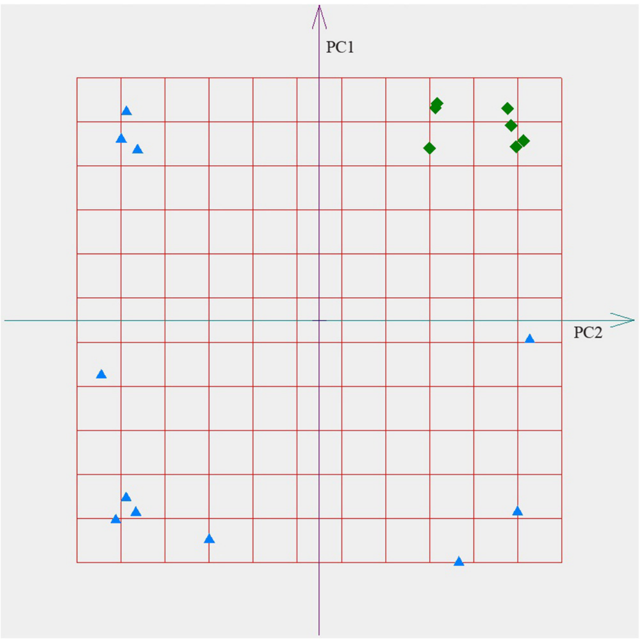

Fig. 3 Spatial distribution of the exhaled air samples from the COPD patients (the diamond icons) and healthy volunteers (the triangle icons). The feature set includes absorption coefficients of the sample in the range of 3.272 to 3.498 μm. The axes correspond to the first (PC1) and the second (PC2) principal components.

The similar results of point estimates of the measured spectra of the exhaled air of COPD patients and healthy volunteers from the control group using gas analyzer LaserBreeze in the most informative subranges from 2.59 to 4.18 μm are shown in Figs. 2–4.

Figures 2–4 show that the methods of IR LPAS and PCA allow separating patients with COPD and healthy nonsmoking volunteers.

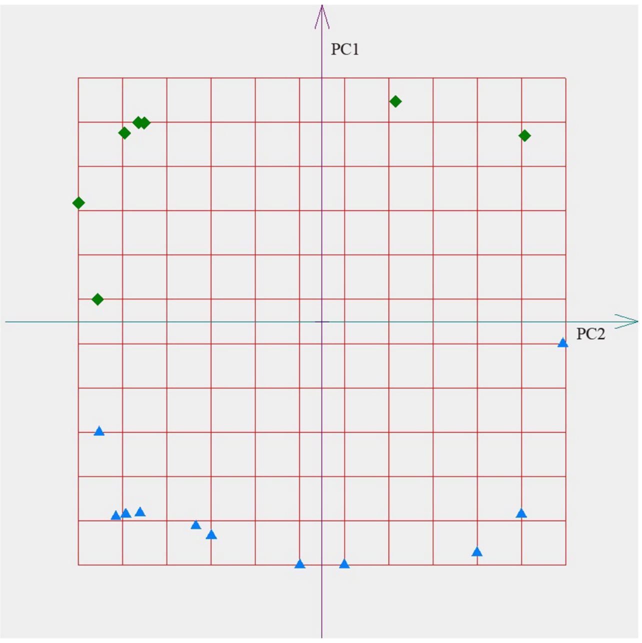

Fig. 4 Spatial distribution of the exhaled air samples from the COPD patients (the diamond icons) and healthy volunteers (the triangle icons). The feature set includes absorption coefficients of the sample in the range of 3.499 to 3.725 μm. The axes are corresponded to the first (PC1) and the second (PC2) principal components.

Table 6 The classification of the absorption spectra of exhaled air of patients with COPD and healthy volunteers by SIMCA in the range of 2.59 to 2.817 μm.

Quantity of samples of the absorption spectra scans in the training stage	Quantity of samples of the absorption spectra scans in the testing stage	Average classification accuracy[a] (%)
6	83	89.46
12	77	96.00
18	71	93.31
24	65	86.15
30	59	71.19

[a]The average value was calculated using 12 different variants of the scans' sets for the training stage.

Taking into account that the VOCs absorption bands correspond to the chosen spectral ranges, we can assume that spatial separation of the target and control groups is probably caused by the difference of hydrocarbon content in the COPD patients' breath and the healthy nonsmoking volunteers' breath. This matches to the results obtained by means of other methods for the analysis of the exhaled air.[2]

For further analysis, the absorption spectrum profile of a breath sample is suggested to be recognized as a "fingerprint" of the medical state of a patient. In order to estimate the possibility of diagnostics based on such "fingerprints," the algorithm of soft*independent modeling of class analogy (SIMCA) was applied. SIMCA classification includes two stages.

The training stage. Each class of objects from the training set is independently modeled using PCA. In this result, the initial data are presented in the cloud form in the space of PCs. The coordinate origin is placed at the center of gravity of the cloud. Each object can be represented as the sum of two vectors: one lying in the cloud (projection) and another perpendicular to the first (residues). The average value (range) and deviation of the lengths of these vectors are the indicators belonging to this class of objects.

The testing stage. The classification procedure is as follows. Each new object is projected onto the built space (cloud). The obtained range and deviation are compared with the critical levels specified in the training stage.

In Table 6, we presented the examples of SIMCA classification using the profiles of the absorption spectra of breath samples in the range of 2.59 to 2.817 μm for various sets of samples for the training and testing stages.

The results in Table 6 show that analysis of the profile of exhaled air absorption spectra in the IR region allows us to separate COPD patients from the control group with a high enough accuracy.

4 Conclusion

We described two types of laser photoacoustic gas analyzers which were developed by Special Technologies Ltd. for medical applications. Laboratory research of the exhaled air of patients with COPD and healthy nonsmoking volunteers was carried out at Siberian State Medical University (Russia) and in the Tomsk State University (Russia). The PCA method was used to select the most informative ranges of the absorption spectra of patients' exhaled air in terms of the separation of the studied groups. It is shown that analysis of the profile of the exhaled air absorption spectrum allows identifying COPD patients in comparison to the control group. The most informative ranges of the absorption spectra of the COPD patients' exhaled air and healthy nonsmoking volunteers' exhaled air are 9.2 to 9.8, 2.59 to 2.817, and 3.272 to 3.725 μm. The presented results are the base for the future construction of the classification rules for the noninvasive express diagnostics methods. There are two ways for classification rules construction. The first is to use the profile of the absorption spectrum of a breath sample as a "fingerprint" of the patient's medical state. Another one consists of two steps: first, to carry out component analysis of breath samples for various groups, then to define the profile of the set of informative VOCs as a "fingerprint" of the patient's medical state.

Acknowledgments

The work was carried out with partial financial support of the FCPIR Contract No. 14.578.21.0082 (ID RFMEFI57814X0082).

References

1. D. Smith and A. Amann, *Breath Analysis for Clinical Diagnosis and Therapeutic Monitoring*, World Scientific, Singapore (2005).
2. D. Smith and A. Amann, *Volatile Biomarkers: Non-Invasive Diagnosis in Physiology and Medicine*, 1st ed., Elsevier, UK (2013).
3. S. A. Kharitonov and P. J. Barnes, "Exhaled markers of pulmonary disease," *Am. J. Respir. Crit. Care Med.* **163**(7), 1693–1722 (2001).
4. M. Yamara, "Exhaled carbon monoxide levels during treatment of acute asthma," *Eur. Respir. J.* **13**, 757–760 (1999).
5. S. Kwiatkowska, "Elevated exhalation of hydrogen peroxide and circulating IL-18 in patients with pulmonary tuberculosis," *Respir. Med.* **101**(3), 574–580 (2007).

6. O. B. Pikas, “Effects of alcoholic beverages on the fatty acid spectrum of the expired air condensate lipids in patients with tuberculosis of the respiratory organs,” *Lik. Sprava* **7–8**, 30–33 (2000).
7. F. K. Tittel, D. Richter, and A. Fried, “Mid-infrared laser applications in spectroscopy,” in *Solid-State Mid-Infrared Laser Sources*, I. T. Sorokina and K. L. Vodopyanov, Eds., pp. 445–516, Springer-Verlag, Berlin, Heidelberg (2003).
8. A. A. Karapuzikov et al., “LaserBreeze gas analyzer for noninvasive diagnostics of air exhaled by patients,” *Phys. Wave Phenom.* **22**(3), 189–196 (2014).
9. V. I. Kozintsev et al., *Laser Photo-Acoustic Analysis of Multicomponent Gaseous Mixture*, Science and Education, Moscow (2003).
10. M. Phillips et al., “Volatile organic compounds in breath as markers of lung cancer: a cross-sectional study,” *Lancet* **353**(9168), 1930–1933 (1999).
11. D. Poli et al., “Exhaled volatile organic compounds in patients with non-small cell lung cancer: cross sectional and nested short-term follow-up study,” *Respir. Res.* **6**(1), 71 (2005).
12. A. L. Pomerantsev and O. Ye Rodionova, “Concept and role of extreme objects in PCA/SIMCA,” *J. Chemom.* **28**(5), 429–438 (2014).
13. A. D. Wilson and M. Baietto, “Advances in electronic-nose technologies developed for biomedical applications,” *Sensors* **11**, 1105–1176 (2011).
14. R. Cangelosi and A. Goriely, “Component retention in principal component analysis with application to cDNA microarray data,” *Biol. Direct* **2**, 21 (2007).
15. A. N. Gorban and A. Y. Zinovyev, “Principal graphs and manifolds,” in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques*, E. S. Olivas et al., Eds., pp. 28–59, Information Science Reference, IGI Global, Hershey, Pennsylvania (2009).

Yury V. Kistenev is the author of more than 120 journal papers, including patents and conference proceedings. His current research interests include application of laser photoacoustic spectroscopy in medicine and biology.

Biographies for the other authors are not available.