Journal of Biomedical Optics

BiomedicalOptics.SPIEDigitalLibrary.org

Exhaled air analysis using wideband wave number tuning range infrared laser photoacoustic spectroscopy

Yury V. Kistenev
Alexey V. Borisov
Dmitry A. Kuzmin
Olga V. Penkova
Nadezhda Y. Kostyukova
Alexey A. Karapuzikov

SPIE.

Yury V. Kistenev, Alexey V. Borisov, Dmitry A. Kuzmin, Olga V. Penkova, Nadezhda Y. Kostyukova, Alexey A. Karapuzikov, "Exhaled air analysis using wideband wave number tuning range infrared laser photoacoustic spectroscopy," *J. Biomed. Opt.* **22**(1), 017002 (2017), doi: 10.1117/1.JBO.22.1.017002.

Journal of Biomedical Optics 22(1), 017002 (January 2017)

Exhaled air analysis using wideband wave number tuning range infrared laser photoacoustic spectroscopy

Yury V. Kistenev,[a,b,*] Alexey V. Borisov,[a,b] Dmitry A. Kuzmin,[a,b] Olga V. Penkova,[a] Nadezhda Y. Kostyukova,[c] and Alexey A. Karapuzikov[c]
[a]National Research Tomsk State University, 36 Lenin Avenue, Tomsk 634050, Russia
[b]Siberian State Medical University, 2 Moscovsky Trakt, Tomsk 634050, Russia
[c]Special Technologies, Ltd., 1/3 Zelenaya Gorka, Novosibirsk 630060, Russia

Abstract. The infrared laser photoacoustic spectroscopy (LPAS) and the pattern-recognition-based approach for noninvasive express diagnostics of pulmonary diseases on the basis of absorption spectra analysis of the patient's exhaled air are presented. The study involved lung cancer patients ($N = 9$), patients with chronic obstructive pulmonary disease ($N = 12$), and a control group of healthy, nonsmoking volunteers ($N = 11$). The analysis of the measured absorption spectra was based at first on reduction of the dimension of the feature space using principal component analysis; thereafter, the dichotomous classification was carried out using the support vector machine. The gas chromatography–mass spectrometry method (GC–MS) was used as the reference. The estimated mean value of the sensitivity of exhaled air sample analysis by the LPAS in dichotomous classification was not less than 90% and specificity was not less than 69%; the analogous results of analysis by GC–MS were 68% and 60%, respectively. Also, the approach to differential diagnostics based on the set of SVM classifiers usage is presented. [DOI: 10.1117/1.JBO.22.1.017002]

Keywords: exhaled air; lung cancer; chronic obstructive pulmonary disease; noninvasive express diagnostics; volatile organic compounds; laser photoacoustic spectroscopy.

Paper 160370PRRR received Jul. 26, 2016; accepted for publication Jan. 4, 2017; published online Jan. 25, 2017.

1 Introduction

The analysis of exhaled air is under investigation as a promising tool for express and noninvasive analysis of biochemical processes in the human body[1] that arise from underlying diseases by providing a detailed picture of specific metabolites that are biomarkers in the exhaled air.[2] The term "biomarkers" was first used in 1989 (Ref. 3) and standardized in 2001, as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention."[4] Control of metabolites in exhaled air produced by biochemical reactions in cells being called "breathomics" provides the ability to predict the specific disease before the appearance of the clinical features. This approach has already been applied for diagnostics of cancer, pulmonary diseases, and infectious diseases.[5]

In addition to nitrogen, oxygen, carbon dioxide, water vapor, and inert gases, exhaled air contains components of endogenous or exogenous origin in the ppbv–pptv range of concentrations. The endogenous compounds include inorganic gases such as NO, CO; volatile organic compounds (VOCs) such as ethane, pentane, acetone, isoprene, acetaldehyde, methanol, ethanol, and other alcohols and alkanes; 2-propanol, sulfur-containing compounds such as dimethylsulfide; methyl, ethyl, mercaptanes, and carbon disulfide; and nitrogen-containing substances such as ammonia and dimethyl/trimethylamine.[1,6]

Single-molecule biomarkers often do not suffice for describing a specific phenotype or endotype. Therefore, molecular biomarker panels are often applied as they can be highly relevant in distinguishing subgroups of patients for targeted interventions. These panels can be derived from complete mapping of molecular mixtures obtained from "omics" technologies and subsequent unbiased statistical pattern recognition.[6]

Exhaled air analysis can be used both as a tool in diagnostics and to reveal specific (patho-) physiological mechanisms. The latter is not of primary importance for diagnostic purposes. Therefore, identification of VOCs is not strictly necessary in a clinical setting, and a "profiling" approach can be used.[7] Chemical analytical techniques provide identification of specific compounds, pattern-recognition-based techniques provide probabilistic discrimination of biomarker profiles. Notably, the latter does not identify individual compounds but is based on probabilistic recognition, which forms the basis for assessing diagnostic accuracy.[7]

The aim of this paper is to reveal the abilities of the infrared (IR) laser photoacoustic spectroscopy (LPAS) and the pattern-recognition-based approach for noninvasive express diagnostics of pulmonary diseases on the basis of absorption spectra analysis of the patient's exhaled breath. The method of gas chromatography–mass spectrometry (GC–MS) was used as the reference.

2 Technical Background

Various analytical methods are used for breathomics.[8] Selected ion flow tube mass spectrometry (SIFT-MS) is based on chemical ionization (ChI) using molecular ions to transfer charge onto

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

1083-3668/2017/$25.00 © 2017 SPIE

the target compound. The ChI approach allows reduced fragmentation of the latter in comparison with many other types of ionization. SIFT-MS provides direct analysis with no sample preconcentration, is suitable for real-time monitoring, and is slightly influenced by humidity. The limit of detection (LOD) of the SIFT-MS Voice200Ultra (Syft Technologies Ltd.) is better than 1 pptv.

Proton transfer reaction mass spectrometry (PTR-MS) is a ChI mass spectrometric technique, which allows the measurement of trace gases as, for example, in exhaled human breath. To increase measuring accuracy, the duration of the measuring process is extended, but for breath-to-breath resolution the time window for measurement should be relatively short. To estimate the LOD, a theoretical model of the measurement process is outlined. According to this, for example, LOD for concentration measurements of the acetone is about 0.2 ppb.[9] The PTR-QMS 300 instrument (IONIKON Analytik GmbH) provides LOD < 300 pptv.[10]

Gas chromatography–mass spectrometry (GC–MS) detection is a "gold standard" in VOCs analysis. For example, LOD for dichloromethane by this method is about 0.1 ppt.[11]

The method of ion mobility spectrometry (IMS) is used to detect substances in very small concentrations, for instance for measurements of background concentrations of pollutants in workplace and environment. A small sample of air containing the suspected substance is periodically taken into the IMS system where a radioactive source ionizes the molecules in the sample. As a result, they drift in an electric field inside the so-called "drift cell." Each type of molecules has a specific drift velocity in the air and may, therefore, be identified. Gas chromatography coupled to ion mobility spectrometer (GC-IMS) by Gesellschaft für analytische Sensorsysteme mbH provides a typical value of LOD near the low ppbv-range.[12]

The devices, consisting of a number of sensors, each of which corresponds to a particular substance, are often called "electronic nose." The example of the "e-nose" is "Cyranose 320," consisting of 32 polymer chemiresistors.[13] The disadvantage of similar sensors is nonspecificity due not only to reaction on a given chemical compound but also to sensibility to nearly all compounds, and slightly more to one chemical family, such as organic solvents, fatty acids, sulfurous gases, etc.[14]

LPAS is one of the most sensitive approaches of laser absorption spectroscopy to gas analysis, especially with the use of coherent radiation sources and intracavity photoacoustic detection.[15] LPAS has a very low detection limit. For example, LPAS gas analyzer with intracavity acoustic cell provides the measurement of ethylene down to 6 pptv.[16,17] Several milliliters of gas sample volume is enough for LPAS analysis. Sample preconcentration is not needed because the photoacoustic signal is proportional to the absorbed volume fraction of laser energy, which can be increased by the power of the used laser source. Therefore, it is preferable to use as it is as high-power a light source as is available.

Light sources that have been used in photoacoustic spectroscopy include broadband infrared radiation sources, that is, black-body radiators and light-emitting diodes; in most cases, various lasers (CO_2, CO, diode, quantum cascade, and Nd:YAG lasers) are used. Another way to use Nd:YAG lasers in LPAS is optical parametric oscillator (OPO) systems as a source of high-power, continuously tunable mid-IR light.[18] OPO systems provide light power of a few 100 mW to more than 1 W in the wavelength range from 2 to 4 μm. OPOs were first used in photoacoustic detection of organic compounds near 3.3 μm at ppm-level and successfully applied later to measure formaldehyde with ppb and ethane with sub-ppb accuracy.[19]

The sensitivity of LPAS is strongly influenced by the construction of the photoacoustic cell. The latter can operate either in a nonresonant mode or as an acoustic resonator. Nonresonant operation means that the light modulation frequency is below the lowest resonance frequency of the cell. In this case, acoustic wave distribution within the cell is almost spatially independent and resonant amplification of the photoacoustic signal is not used. When the exciting light is modulated at a resonance frequency of the cell, the generated photoacoustic signal is amplified proportionally to the quality factor (Q-factor) of the acoustic resonance. Q-factors can be up to several hundreds.[15]

The most frequently used types of resonant LPAS detectors are based on Helmholtz resonators, one-dimensional cylindrical resonators, and cavity resonators.[15,20] The Groupe de Spectrométrie Moléculaire et Atmosphérique (Reims, France) and the Institute of Atmospheric Optics (Tomsk, Russia) have developed a photoacoustic sensor based on a double differential Helmholtz resonator (DHR) for infrared gas detection.[20–22] The double DHR uses two identical DHR configurations, which can significantly eliminate the in-phase external acoustic noise at atmospheric pressure and flow mode.[18]

Nonlinear effects in OPO is one of the most widespread ways to generate tunable coherent radiation in the wide spectral range. We developed the LaserBreeze gas analyzer based on an LPAS method and OPO with a tuning range from 2.5 to 10.7 μm.[23]

The experimental set-up of the LaserBreeze gas analyzer is shown in Fig. 1. The laser source includes two OPOs. The first one is based on fan-out periodically poled lithium niobate structure (PPLN), which provides wavelength tuning in the spectral range from 2.5 to 4.5 μm. The second OPO is based on mercury thiogallate crystals $HgGa_2S_4$ (HGS) and has a wavelength tuning range from 4.45 to 10.7 μm. Both OPO were pumped by a Nd:YLF laser. The switching between two OPO is realized by a motorized translation stage. The linewidth of laser radiation is about 3 to 4 cm^{-1}. It is enough for a pattern-recognition-based approach. Resolution of wavelength scanning is around 7 nm/s for OPO based on PPLN structure. This value for OPO based on HGS crystal due to its mechanism of wavelength tuning is varied over spectral range, but the values are practically the same. The total time of the absorption spectrum registration in the whole spectral range is about 10 min. The photoacoustic

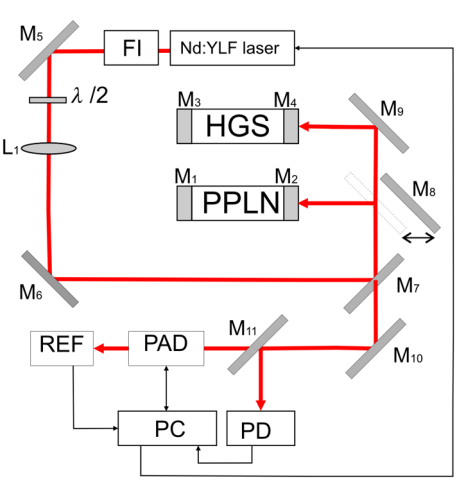

Fig. 1 Experimental setup of the LaserBreeze.

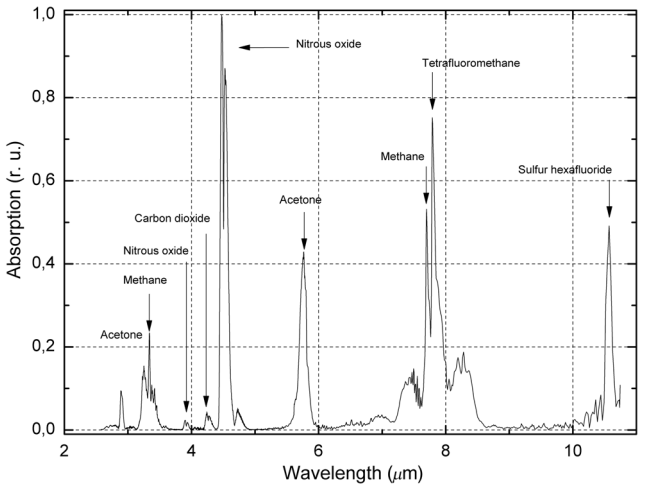

Fig. 2 Absorption spectra of gas mixture in the reference cell in the spectral range from 2500 to 10,700 nm.

detector (PAD) is based on double channel Helmholtz resonator with Q-factor ~40 and fundamental resonance frequency ~1700 Hz. Data from the pyroelectric detector (PD) are used to normalize the PAD signal relative to the laser radiation power. The thermostating at the temperature $40°C \pm 0.2°C$ was applied to avoid temperature drift of the OPO parameters and water vapor condensation on the PAD walls.

To provide wavelength calibration, we use the reference cell (REF) filled with a gas mixture with a known composition of compounds having strong absorption lines in known wavelengths within the LaserBreeze gas analyzer tuning range. Absorption spectrum of the gas mixture in the reference cell is shown in Fig. 2.

The other designation in Fig. 1 are: FI is the Faraday isolator, M_i are the mirrors, PC is the personal computer, and $\lambda/2$ is the halfwave plate.

In the case of a smooth absorption spectrum with no distinct peaks of absorption of spectral bands of the measuring components of a gas mixture, a method based on Bayesian estimate of the solution of the inverse spectroscopy task allows the determination of the gas concentration.[24] The LaserBreeze gas analyzer allows the detection of more than 20 molecular biomarkers that have absorption lines in the mentioned spectral range, including acetone (C_3H_6O), acetylene (C_2H_2), ammonia (NH_3), butane (C_4H_{10}), carbon dioxide (CO_2), 13 isotope of carbon dioxide ($^{13}CO_2$), carbon monoxide (CO), ethane (C_2H_6), ethanol (C_2H_5OH), ethyl acetate ($C_4H_8O_2$), ethylene (C_2H_4), formaldehyde (CH_2O), methane (CH_4), methanol (CH_3OH), nitrogen dioxide (NO_2), nitrogen oxide (NO), nitrous oxide (N_2O), pentane (C_5H1_2), propane (C_3H_8), and sulfur dioxide (SO_2). Relative error in determining of VOC concentrations is not more than 30%.

The necessary volume of the studied sample is not more than 50 cm^3, and the concentration sensitivity of the LaserBreeze gas analyzer is not worse than 1×10^{-3} ppm.

A procedure of sensitivity estimation was described in Ref. 22. PAD was preliminarily cleared by the pumping of N_2. After that, the device was switched on. The measurements of noise signal value U_N were continued for 3 min. The average value $\langle U_N \rangle$ and standard deviation δU_N were calculated. Then, PAD was filled by a calibration gas mixture including tested gas with known concentrations n and nitrogen (N_2). The concentration of tested gas was chosen to provide a useful signal value U_S over U_N in 2 to 3 times. The measurements procedure was the same as for noise level one. The following equation was used to calculate signal/noise value (S/N):

$$S/N = \frac{\langle U_s \rangle}{\langle U_N \rangle + \delta U_N},$$

where $\langle U_s \rangle$ is the average value of useful signal. The sensitivity n_o was determined by the following equation:

$$n_o = \frac{n}{S/N}. \tag{1}$$

3 Data Preprocessing and Analysis

One of the key steps in the biomarkers analysis involves evaluation of latent dependencies in the variables data using reliable methods. The methods often are referred to as chemometrics.

The first step in chemometrics data analysis usually consists of separation of informative variables and reduction of the dimension of the feature space. This can be provided by multivariate unsupervised methods such as principal component analysis (PCA), factor analysis, k-means clustering, or hierarchical cluster analysis.[25]

The basic idea of PCA is to find the reduced number of new variables, termed the principal components, that are enough for the recovery of the initial variables, possibly with insignificant errors. The mathematical background of PCA consists of decomposition of initial experimental data from a two-dimensional matrix X ($I \times J$) in the form of a matrix product[26]

$$X = T \cdot P^t + E, \tag{2}$$

where T, P, and E are the scores, loadings, and residuals matrixes, respectively. The loadings matrix contains weight coefficients that characterize the contribution of features to a principal component. The scores matrix contains coordinates of the samples in the space of the principal components.

Breathomics data frequently show nonlinear patterns in the feature space, and these problems are well handled using nonlinear methods.[27] Nonlinear techniques, particularly kernel methods, are more powerful in predicting accuracy and discrimination.[28] The support vector machine (SVM) is the most frequently used kernel method.

SVM binary classification is based on building up the maximum-width stripe that spatially separates groups under study. The algorithm is based on scalar product analysis of the feature vectors. When the building of such a stripe is impossible, the kernel transform can help to provide classification that is based on analysis of the scalar product of the feature vectors functions. The application of SVM to the problem of data classification is by a training set with objects that belong to one of the two classes; each new object is assigned to one of these classes. The problem may be defined as follows:

$$(x_1, y_1), \ldots, (x_m, y_m) \in \mathbf{X} \times \{\pm 1\}, \tag{3}$$

where $\mathbf{X}$ is a nonempty set; m is the number of objects in the training set; y_i are called labels, and x_i are the objects under classification. Each classified object is a vector in n-dimensional space.

Thus, the task of some classifier rule building is

$$a(x) = \text{sign}\left(\sum_{j=1}^{n} w_j \cdot x^j - b\right) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle - b), \quad (4)$$

where operation $\langle \mathbf{w}, \mathbf{x} \rangle$ defines the scalar product of vectors, and the vector $w = (w_1, w_2, \ldots, w_n) \in \mathbb{R}^n$ and scalar threshold $b \in \mathbb{R}$ are the algorithm parameters.

The SVM method includes a training phase; consequently, the experimental data set should be separated into teaching and testing subsets. The separation procedure essentially influences the robustness of the classification. This can include both a cross validation and an external validation to avoid discrepancy. In the n-fold cross validation, the dataset is randomly divided into n subsets of equal size; after that, $(n-1)$ subsets are used for training and the remaining subset is used for the examination of classification quality. This procedure should be repeated until all n subsets have been used as the test set.[29,30] The limit case of this algorithm is "leave one out cross validation," which corresponds to n being equal to the experimental data set size. In the external validation, a new dataset obtained by repetition of the measurements with the same population is used.[31,32]

4 Results and Discussion

The experimental part of the research was carried out according to the principles of good clinical practices. Protocol of the research was approved by the Ethic Committee of the Siberian State Medical University (Tomsk, Russia), Ref. No. 2882 at 24.11.2011. All participants were preliminary informed about details of the research and signed an "informed agreement" on the actions carried out. The interaction with the patients was limited by the sampling of a part of exhaled air into a disposable container.

The sampling procedure occurs before eating or 2 h after. Prior to sampling, participants rinsed their mouths with running water without any special cleaning of the oral cavity. Then, participants did some calm breaths through a sterile plastic tube into the sample container. The "dead volume" was exhaled outside the sample container.

The study involved three groups: patients with bronchopulmonary diseases including lung cancer (LC) ($N = 9$); patients with chronic obstructive pulmonary disease (COPD) ($N = 12$); and a control group of healthy nonsmoking volunteers ($N = 11$). All patients had been treated or diagnosed in specialized units of medical institutions, so the diagnosis of every patient had been verified and thoroughly tested by instrumental methods. All patients with severe comorbidities, with chronicity of the pathological processes, or an unconfirmed clinical diagnosis were excluded from the study.

All patients with COPD were men in the Pulmonological Division of the Regional State Autonomous Institution of Public Health Municipal Clinical Hospital No. 3 (Tomsk, Russia), with an average age of 67.8 ± 9.7 years; 10 of 12 of them were smokers with average smoking of 42 ± 13 years. The details are shown in Table 1.

All LC patients were men in the Thoraco-Abdominal Division of the Federal State Budget Scientific Institution Tomsk National Research Center of the Russian Academy of Medical Sciences (Tomsk, Russia), with an average age of 61.5 ± 4.8 years; 8 of 9 patients were smokers with average smoking of 44.9 ± 8.2 years. The details are shown in Table 2.

The control group consisted of nominally healthy males with an average age of 21.5 ± 1.6 years. Exclusion criteria were the presence of "smoking" in their anamnesis vitae and the presence of diseases of the bronchopulmonary, cardiovascular, digestive, endocrine, reproductive, and urinary organ systems in the chronic form, as well as in the acute form during the 3 weeks prior to sampling.

Table 1 Information about the group of patients with COPD.

Patient ID	Age (years)	Primary diagnosis	Complication	Length of smoking (years)
C1	53	COPD stage II, exacerbation	No	40
C2	70	COPD stage III, exacerbation	Chronic pulmonary heart, compensation	40
C3	71	COPD stage IV, exacerbation	Chronic pulmonary heart, compensation	45
C4	63	COPD stage IV, exacerbation	Chronic respiratory failure I, chronic pulmonary heart, compensation	50
C5	84	COPD stage II, exacerbation	No	No
C6	71	COPD stage III, exacerbation	Chronic pulmonary heart, compensation	50
C7	86	COPD stage II, exacerbation	No	60
C8	66	COPD stage III, exacerbation	No	20
C9	66	COPD stage I, exacerbation	No	20
C10	63	COPD stage IV, exacerbation	Chronic pulmonary heart, compensation	50
C11	65	COPD stage IV, exacerbation	Chronic pulmonary heart, compensation	45
C12	56	COPD stage II, exacerbation	No	No

Table 2 Information about the group of patients with lung cancer.

Patient ID	Age (years)	Primary diagnosis	TNM classification	Length of smoking (years)
L1	61	Peripheral cancer, upper lobe of left lung	$T_4N_2M_0$, IIIB	41
L2	60	Central cancer, upper lobe bronchus on right	$T_4N_1M_0$, IIIA	47
L3	60	Central cancer, lower lobar bronchus on right	$T_2N_2M_0$, IIIA	40
L4	62	Central cancer, intermediate bronchus on right	$T_3N_2M_0$, IIIA	45
L5	65	Peripheral cancer, lower lobe of left lung with spread on chest wall and upper lobe	$T_3N_xM_0$, IIB	50
L6	59	Central cancer, bottom lobar bronchus on left with spread on pulmonary vein	$T_3N_xM_0$, IIB	35
L7	68	Peripheral cancer, upper lobe of left lung with spread on interlobar pleura, metastases of lymph nodes in aortic window	$T_3N_2M_0$, IIIA	35
L8	67	Central cancer, upper lobe bronchus on right with spread on main bronchus, trachea, carina	$T_4N_xM_0$, IIIA	46
L9	52	Central cancer, lower lobar bronchus on left with extensive local spread	$T_4N_3M_0$, IIIB	No

Exhaled breath samples (EBS) were collected in disposable plastic containers (syringe) with a volume of 150 ml and analyzed using the LaserBreeze gas analyzer. Additionally, EBS were collected in the Bio-VOC breath sampler with Supelco solid phase microextraction fiber holder 57330U. The extraction time was 30 min. All measurements were carried out at room temperature (variations were 20°C to 25°C) and humidity (50% to 60%).

The EBS from the Bio-VOC breath sampler were analyzed by gas chromatography Finnigan Trace GC with MS detector Finnigan Trace DSQ (GC–MS). Processing of the data is produced in Qual Browser of Xcalibur software. For identification of VOCs, substances spectra obtained are compared with the substances spectra from the NIST MS Search 2.0 library. After the VOCs were identified, the area of chromatographic peaks was estimated manually on Xcalibur software as a concentration parameter of identified VOCs in EBS.

To validate the suitability of the Bio-VOC breath sampler and plastic containers for sampling of the EBS, we filled both containers with nitrogen of 99% purity and analyzed the content by GC–MS technique. The measured chromatograms had no peaks, which indicate that the used samplers do not contribute any errors in analysis.

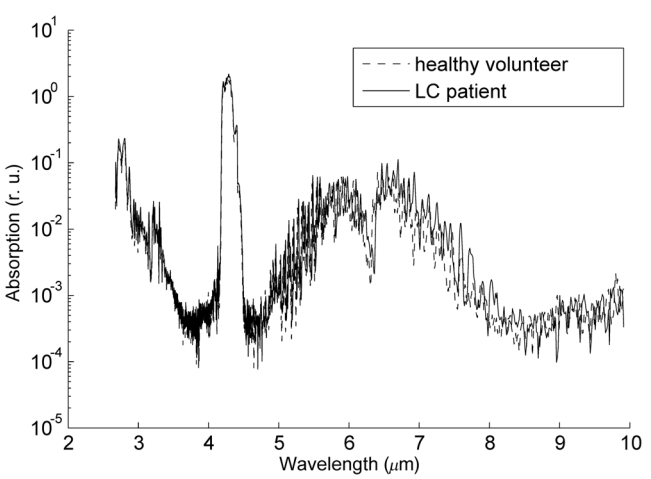

Fig. 3 An example of measured absorption spectra of EBS from LC patient and healthy volunteer in the spectral range from 2600 to 10,000 nm.

An example of measured by the LaserBreeze gas analyzer absorption spectra of EBS from an LC patient and a healthy volunteer is presented in Fig. 3.

In the comparative analysis of slightly different feature vectors in high-dimensional feature space, there is a known problem of the homogeneity (weak visibility) of the similar vectors.[32] To overcome this problem, we provided a two-step analysis of the measured spectra. First, the selection of informative features and reduction of the dimension of the feature space was realized using PCA preprocessing; thereafter, the classification was carried out using SVM. In contrast to standard approaches of PCA-SVM usage, we carried out the optimization procedure used for both classification principal components and SVM kernels and kernel parameters. At the latter step, we used the polynomial kernel, multilayer perceptron kernel, and Gaussian radial basis function.[26]

The teaching and testing sets were produced by splitting the initial data into a specific proportion. The random forming of teaching and testing sets was repeated 50 times, and the results were averaged. The results of dichotomous classification of EBS absorption spectra measured by the LaserBreeze gas analyzer are presented in Table 3.

Figure 4 shows the dichotomous classification of COPD-LC patients using the multilayer perceptron kernel. Round markers correspond to the reference vectors, cross-markers correspond to the COPD patients, and triangles mark LC patients.

The profile of metabolites analyzed in EBS by the GC–MS method includes methanol, ethanol, acetonitrile, acetone, methylene chloride, pentane, ethylacetate, hexane, benzene, propyleneoxide chloride, n-ethylformamide, octane, toluene, butylacetate, chlorobenzene, o-xylene, decane, and chloroform. The results of the dichotomous classification of these profiles by a PCA-SVM combination technique as described above are presented in Table 4. Here, the training set consisted of five patterns for each group.

Comparison of the results presented in Tables 3 and 4 shows that, in our case, the classification results obtained by LPAS are more promising than the results obtained by GC–MS.

Table 3 SVM classification of the testing set of EBS absorption spectra measured by the LaserBreeze gas analyzer for the groups under study (patients with lung cancer, COPD, and healthy volunteers).

Dichotomous classification	SVM kernel	Kernel parameters	Sensitivity		Specificity	
			Mean	Dispersion	Mean	Dispersion
COPD–LC	Gaussian radial basis function	1.1953	0.9258	0.0009	0.7790	0.0584
LC–healthy volunteers	Gaussian radial basis function	0.0832	0.9267	0.0102	0.9191	0.0039
COPD–healthy volunteers	Multilayer perceptron	5.0241 and 24.3958	0.9027	0.0473	0.6894	0.0303

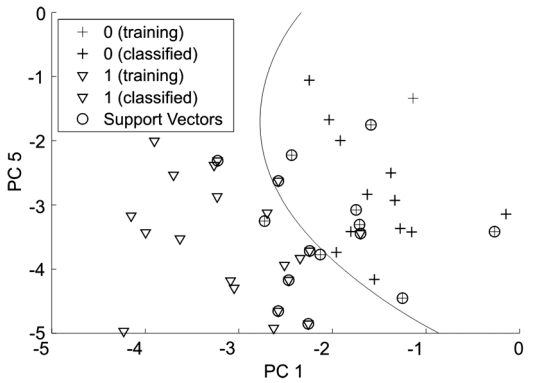

Fig. 4 Classification of EBS absorption spectra from COPD and LC patients in the space of the principal components using SVM with the multilayer perceptron kernel. The projection on the plane of the first and fifth principal components is shown.

The above mentioned results allow one to construct the rules of differential diagnostics based on the set of SVM classifiers usage. There are several approaches to solve this problem using binary classifiers.[33] According to the "One-vs-All" method, we had to construct N-independent binary classifiers, so the every classifier will separate a specific class feature vectors from all other class's feature vectors.[34] According to the "One-vs-One" (also known as "All-vs-All") method, we had to construct $N(N-1)$ independent binary classifiers, each of which will separate i'th class feature vectors from j'th class feature vectors.[35] The latter method was shown to provide the better results.

The results of differential diagnostics based on EBS analysis by LPAS and three SVM dichotomous classifiers from Table 3 and the "One-vs-One" method are presented in Table 5. The estimations were carried out using a merged testing set that included LC, COPD patients, and healthy volunteers, as is shown in Table 5.

Table 5 Differential diagnostics based on the set of SVM classifiers usage.

Group	Quantity of the feature vectors in the testing set	Diagnosis		
		Set right	Set wrong	Did not set
LC	8	8	0	0
COPD	12	10	2	0
Healthy volunteers	29	26	1	3

The feature vector of a representative from the testing set was analyzed by every classifier from Table 3. The differential diagnostics rule was based on the result that was selected more times. Diagnosis did not set if all possible results of classification (LC–COPD–healthy) for definite representative from the testing set met the same number of times.

5 Conclusion

EBS analysis is a promising tool for express and noninvasive analysis of biochemical processes in the human body and diagnosis of various diseases. In other words, a similar technique is useful for identifying specific metabolites in the EBS or for discrimination of metabolites–biomarkers profiles using pattern-recognition-based methods of data analysis. We used IR LPAS and GC–MS methods to provide spectral analysis of EBS. The analysis of measured spectra was based first on reduction of the dimension of the feature space using PCA; thereafter, the dichotomous classification was carried out using a SVM. The estimated average sensitivity of EBS analysis by the LPAS in dichotomous classification was not worse than 90%,

Table 4 SVM classification of the testing set of EBS absorption spectra measured by GC–MS for the groups under study (patients with lung cancer, COPD, and healthy volunteers).

Dichotomous classification	SVM kernel	Kernel parameters	Sensitivity		Specificity	
			Mean	Dispersion	Mean	Dispersion
COPD-LC	Polynomial	4	0.8800	0.0320	0.6400	0.0680
LC-healthy volunteers	Gaussian radial basis function	0.0250	0.8241	0.0043	0.8875	0.0018
COPD-healthy volunteers	Multilayer perceptron	5 and 0.7	0.6800	0.0520	0.6000	0.1400

the average specificity was not worse than 69%, and the analogous results of analysis by GC–MS were 68% and 60%, respectively.

The results obtained in this study show high potential for the application of LPAS spectral analysis of the exhaled air samples in combination with the pattern-recognition-based approach for noninvasive screening tests of pulmonary diseases. The future steps in bringing this technology to clinics should include design of cost-effective and informative measurement devices, for example, specialized medical purpose LPAS equipment without unnecessary abilities and simple to use, accumulation of spectral information about exhaled air samples of patients with a confirmed diagnosis, and finding effective methods of data analysis and classification.

Disclosures

Alexey A. Karapuzikov has a financial interest in Special Technologies, Ltd., which, however, did not provide financial support for this work. Except for this, no conflicts of interest, financial or otherwise, are declared by the authors.

Acknowledgments

The work was carried out with the partial financial support of the FCPIR contract No. 14.578.21.0082 (ID RFMEFI57814X0082). The authors thank Jean Kollantai, Tomsk State University, for style review.

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, Austria (2013).
3. X. Ping, "Evaluation of repeated biomarkers: non-parametric comparison of areas under the receiver operating curve between correlated groups using an optimal weighting scheme," Graduate Theses and Dissertations (2012).http://scholarcommons.usf.edu/etd/4261.
4. Biomarkers Definitions Working Group, "Biomarkers and surrogate endpoints: preferred definitions and conceptual framework," *Clin. Pharmacol. Ther.* **69**(3), 89–95 (2001).
5. A. W. Boots et al., "Exhaled molecular fingerprinting in diagnosis and monitoring: validating volatile promises," *Trends Mol. Med.* **21**(10), 633–644 (2015).
6. S. Kwiatkowska, "Elevated exhalation of hydrogen peroxide and circulating IL-18 in patients with pulmonary tuberculosis," *Respir. Med.* **101** (3), 574–580 (2007).
7. M. P. van der Schee et al., "Breathomics in lung disease," *Chest* **147**(1), 224–231 (2015).
8. C. Lourenço and C. Turner, "Breath analysis in disease diagnosis: methodological considerations and applications," *Metabolites.* **4**, 465–498 (2014).
9. A. Amann et al., "Model based determination of detection limits for proton transfer reaction mass spectrometer," *Meas. Sci. Rev.* **10**(6), 180–188 (2010).
10. W. Lindinger, A. Hansel, and A. Jordan, "On-line monitoring of volatile organic compounds at pptv levels by means of proton-transfer-reaction mass spectrometry (PTR-MS) medical applications, food control and environmental research," *Int. J. Mass Spectrom. Ion Processes* **173**(3), 191–241 (1998).
11. F. Obersteiner and H. A. Bönisch, "Engel An automated gas chromatography time-of-flight mass spectrometry instrument for the quantitative analysis of halocarbons in air," *Atmos. Meas. Tech.* **9**, 179–194 (2016).
12. www.gas-dortmund.de.
13. M. P. Fernandes, S. Venkatesh, and B. G. Sudarshan, "Early detection of lung cancer using nano-nose—a review," *Open Biomed. Eng. J.* **9**, 228–233 (2015).
14. M. Kuske, A.-C. Romain, and J. Nicolas, "Microbial volatile organic compounds as indicators of fungi. Can an electronic nose detect fungi in indoor environments?" *Build. Environ.* **40** (6), 824–831 (2005).
15. A. Miklós, P. Hess, and Z. Bozóki, "Application of acoustic resonators in photoacoustic trace gas analysis," *Rev. Sci. Instrum.* **72** (4), 1937–1955 (2001).
16. J. A. de Gouw et al., "Airborne measurements of ethene from industrial sources using laser photo-acoustic spectroscopy," *Environ. Sci. Technol.* **43** (7), 2437–2442 (2009).
17. F. G. C. Bijnen, J. Reuss, and F. J. M. Harren, "Geometrical optimization of a longitudinal resonant photoacoustic cell for sensitive and fast trace gas detection," *Rev. Sci. Instrum.* **67**(8), 2914–2923 (1996).
18. Z. Bozóki, A. Pogány, and G. Szabó, "Photoacoustic instruments for practical applications: present, potentials, and future challenges," *Appl. Spec. Rev.* **46**, 1–37 (2011).
19. J. Li, W. Chen, and B. Yu, "Recent progress on infrared photoacoustic spectroscopy techniques," *Appl. Spectr. Rev.* **46**, 440–471 (2011).
20. V. Zéninari et al., "Photoacoustic detection of methane in large concentrations with a Helmholtz sensor: simulation and experimentation," *Int. J. Thermophys.* **37**(1), 1–11 (2016).
21. V. Zéninari et al., "Helmholtz resonant photoacoustic cell for spectroscopy of weakly absorbing gases and gas analysis," *Atmos. Oceanic opt.* **12**(10), 928–940 (1999).
22. C.-M. Lee et al., "High-sensitivity laser photoacoustic leak detector," *Opt. Eng.* **46**(6), 065002 (2007).
23. A. I. Karapuzikov et al., "LaserBreeze gas analyzer for noninvasive diagnostics of air exhaled by patients," *Phys. Wave Phen.* **22**(3), 189–196 (2014).
24. L. N. Eremenko, V. I. Kozintsev, and V. A. Gorodnichev, "Method of Bayesian estimates in the problem of laser gas analysis," *Russ. Phys. J.* **51**(9), 912–918 (2008).
25. A. Kotłowska, "Application of chemometric techniques in search of clinically applicable biomarkers of disease," *Drug Dev. Res.* **75**, 283–290 (2014).
26. L. Pomerantsev and O. Y. Rodionova, "Concept and role of extreme objects in PCA/SIMCA," *J. Chemom.* **28**(5), 429–438 (2014).
27. G. R. G. Lanckriet et al., "Learning the kernel matrix with semidefinite programming," *J. Mach. Learn. Res.* **5**, 27–72 (2004).
28. J. Pereira et al., "Breath analysis as a potential and non-invasive frontier in disease diagnosis: an overview," *Metabolites* **5**, 3–55 (2015).
29. R. R. Picard and R. D. Cook, "Cross-validation of regression models," *J. Am. Stat. Assoc.* **79**(387), 575–583 (1984).
30. J. Xia et al., "Translational biomarker discovery in clinical metabolomics: an introductory tutorial," *Metabolomics.* **9**, 280–299 (2013).
31. A. Krilaviciute et al., "Detection of cancer through exhaled breath: a systematic review," *Oncotarget* **6** (36) (2015).
32. M. B. Shapiro and R. B. Marimont, "Nearest neighbour searches and the curse of dimensionality," *IMA J. Appl. Math.* **24**, 59–70 (1979).
33. M. Aly, "Survey on multiclass classification methods," Technical report, pp. 1–9, California Institute of Technology, Pasadena, California (2005).
34. X. Zhao, S. Guan, and K. L. Man, "An output grouping based approach to multiclass classification using support vector machines," *Adv. Multimedia Ubiquitous Eng.* **393**, 389–395 (2016).
35. J. Milgram, M. Cheriet, and R. Sabourin, "'One against one' or 'one against all': which one is better for handwriting recognition with SVMs?" in *10th Int. Workshop on Frontiers in Handwriting Recognition* (2006).

Yury V. Kistenev is a professor, deputy vice rector for Research of TSU, and he 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.

Alexey V. Borisov, PhD, is an associate professor at TSU. His areas of scientific interests are biomedicine, optics, numerical analysis, and mathematical physics.

Dmitry A. Kuzmin is a junior researcher of SSMU, and he is the author of more than 20 research papers. The present research

interests include gas analysis, laser IR photoacoustic spectroscopy, data mining, and chemometrics.

Olga V. Penkova is a junior researcher of TSU, and she is specialist in quantitative gas chromatographic analysis.

Nadezhda Y. Kostyukova is a engineer of Special Technologies, Ltd., is the author of more than 20 journal papers, including conference proceedings. Her research interests include the development of parametric conversion devices in the mid-IR spectral range.

Alexey A. Karapuzikov is the director of Special Technologies, Ltd. His research interests include development of IR laser sources and laser photoacoustic spectroscopy systems.