Breath air analysis using wide-band tuning range IR laser photoacoustic spectroscopy and machine learning

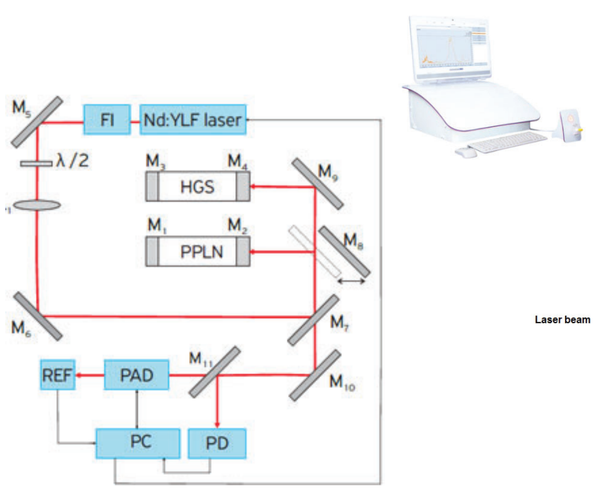
Yury V. Kistenev[1,2,*], Alexey V. Borisov[1,2], Dmitry A. Kuzmin[2], Denis A.Vrazhnov[3], Olga V.Penkova[1]

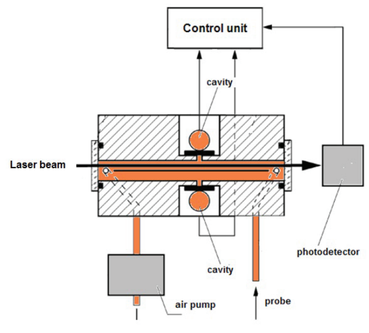
[1]National Research Tomsk State University, 36 Lenin av., Tomsk, Russia;
[2]Siberian State Medical University, 2 Moscowski Trakt St., Tomsk, Russia;
[3]Institute of Strength Physics and Materials Science SB RAS, Tomsk, Russia
* **yv.kistenev@gmail.com**

Abstract
The infrared laser photoacoustic spectroscopy (LPAS) abilities and the pattern-recognition-based approach for non-invasive express diagnostics of pulmonary diseases based on absorption spectra analysis of the patient's breath air are discussed. The study was involved with lung cancer patients (N=30), patients with chronic obstructive pulmonary disease (N= 40), pneumonia (N= 40), and a control group of 130 healthy non-smoking volunteers. The analysis of measured spectra was based, at first, on the reduction of the dimension of the feature space using Principal Component Analysis. Then, the multi-group One-Vs-One classification has been carried out using Support Vector Machine. The method of gas-chromatography-mass-spectrometry (GC-MS) was used as a reference one. The estimated sensitivity of breath air samples analysis with the LPAS in dichotomous classification was not worse than 86%, and the specificity was not worse than 83%. The analogous results in dichotomous classification with GC-MS were 68% and 60%, correspondingly.

Instrument base

Laser OPO optical-acoustic gas analyzer "LaserBreeze"

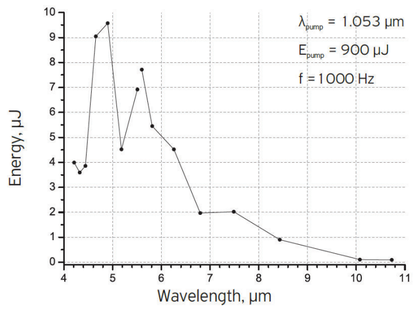


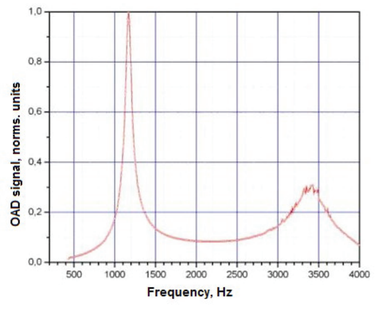


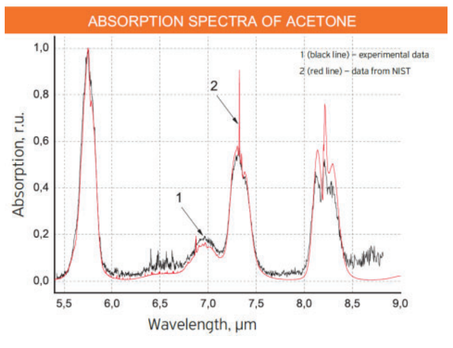
Main parameters	Value
Concentration sensitivity (S/N)	No worse than 1×10^{-3}ppm
Number of detected molecular biomarkers	No less than 20
Scanning range of OPO radiation	2.5 - 10.7 μm

Gas Chromatograph Finnigan Trace GC/Finnigan Trace DSQ

Technical parameters	Value
The mass range	1 – 1050 a.e.m.
Scanning speed	not less than 10000 a.e.m. /sec
Detection limit	not less than 2*10-12 g/sec
The linear range	not less than10+7.







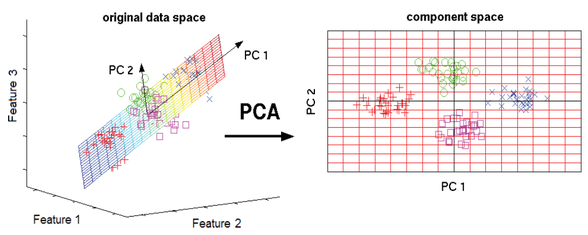
BioVOC sampler

Groups under study and analytical methods

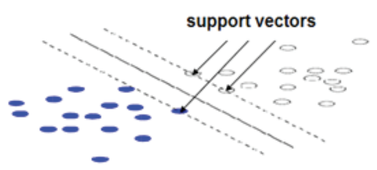
The groups under study

The group	Lung Cancer	COPD/ Pneumonia	Healthy volunteers
Number of participants	30	40/40	130
Average age, years	56.4	53.1	24.7

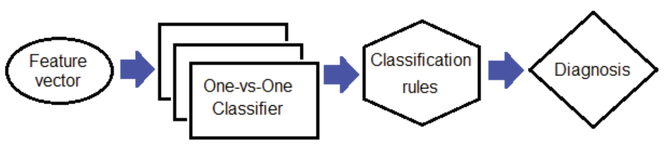
Informative features extraction (Principal component analysis)



Supervised Learning classification (Support Vector Machine)



Differential diagnosis based on the set of SVM One-vs-One classifiers



Results

Laser optical-acoustic spectroscopy + machine learning

SVM binary classification of the testing set of breath air absorption spectra

Groups	Sensitivity		Specificity	
	Mean	Dispersion	Mean	Dispersion
LC- Pneumonia	0,96	0,0014	0,93	0,0012
LC-COPD	0,98	0,0003	0,94	0,0007
LC-Healthy volunteers	0,96	0,0011	0,90	0,0013
COPD-Pneumonia	0,95	0,0016	0,95	0,0012
COPD - Healthy volunteers	0,86	0,0022	0,83	0,0020
Pneumonia - Healthy volunteers	0,96	0,0009	0,92	0,0019

Gas-chromatography-Mass-spectrometry + machine learning

SVM binary classification of the testing set of breath air absorption spectra

Groups	**Sensitivity**		**Specificity**	
	Mean	Dispersion	Mean	Dispersion
LC-COPD	0,88	0,0054	0,83	0,0037
LC-Healthy volunteers	0,95	0,0053	0,92	0,0057
COPD-Healthy volunteers	0.68	0.052	0.60	0.14

Differential diagnosis based on the set of SVM One-vs-One classifiers

Groups	Diagnosis					
	Right diagnosis		Wrong diagnosis		Not set	
	Mean	Dispersion	Mean	Dispersion	Mean	Dispersion
LC	0,9565	0,0013	0,0341	0,0011	0,0094	0,0013
COPD	0,8112	0,0091	0,0981	0,0082	0,0907	0,0047
Pneumonia	0,8412	0,0048	0,0991	0,0032	0,0597	0,0025
Healthy volunteers	0,8946	0,0038	0,0901	0,0024	0,0153	0,0018

Conclusion

The "profiling" approach, based on the set of markers control or profile of the absorption spectrum of breath sample as a "fingerprint" of the state, is presented. We used the IR LPAS method to measure the absorption spectra of exhaled air samples. The analysis of measured spectra was based first on the reduction of the dimension of the feature space using PCA; thereafter, the dichotomous classification was carried out using a SVM. The SVM method provides binary classification, i.e., it can separate objects only into two classes. For purposes of differential diagnostics, it is necessary to construct the classification rules on several classes. To solve this problem, we used the "One-vs-One" approach. The accuracy of classification by the "One-vs-One" method based on spectral analysis of patients' exhaled air is high enough for using in routine practices, especially for screening tests.

The list of publications of our group

This work was partially supposed by the Federal Target Program for Research and Development, Contract No. 14.578.21.0082.