DOI 10.1007/s10527-020-09942-0
*Biomedical Engineering, Vol. 53, No. 5, January, 2020, pp. 355-359. Translated from Meditsinskaya Tekhnika, Vol. 53, No. 5, Sep.-Oct., 2019, pp. 42-45. Original article submitted March 22, 2019.*

# Methods for Processing Mass Spectrometry Signals from Exhaled Gases for Medical Diagnosis

**V. V. Manoilov*, L. V. Novikov, I. V. Zarutskii, A. G. Kuz'min, and Yu. A. Titov**

*The composition of gases exhaled by humans can be analyzed using small-scale quadrupole mass spectrometers. The analysis results can be used for diagnosis of the state of health. Mathematical processing of mass spectrometry signals allows mass spectra to be classified into healthy or sick, and also into groups of patients. Algorithms for primary data processing and classification based on multidimensional statistical methods, i.e., discriminant and cluster analysis, are described.*

## Introduction

Analysis of the composition of exhaled air is a major approach in noninvasive medicine. Diagnosis based on analysis of exhaled gases has a number of advantages over conventional laboratory methods. Analysis of gas mixes is safe for staff as there is no use of chemicals or biological fluids. It is also relatively cheap, rapid, and allows detection of volatile components in exhaled air at the level of trace substance concentrations in real time. This article presents algorithms for processing the mass spectra of exhaled gases obtained using an MS7-200 quadrupole mass spectrometer with electron ionization and direct capillary sample loading [1-3]. The gas to be analyzed is injected into the ionization chamber of the electron-impact ion source at atmospheric pressure via the capillary input. The resulting ions are introduced to the quadrupole-type mass analyzer. The mass spectrometric signals produced are processed using special software and are compared with spectra in a library of standard mass spectra for identification of individual spectral components and assessment of their concentrations. The capillary sample loading system of the mass spectrometer, heated to 50°C, allows analysis to be performed at a distance of up to 5 m from the apparatus. Analysis uses up to 1 mL of sample per minute. The vacuum system is based on diffusion or turbomolecular pumps.

Institute for Analytical Instrumentation, Russian Academy of Sciences, St. Petersburg, Russia; E-mail: manoilov_vv@mail.ru
* To whom correspondence should be addressed.

The aim of the present work was to analyze the potentials of algorithms for processing information to classify patients into healthy and sick groups.

Conventional methods for analysis of mass spectrometric data generally include preliminary (primary) processing for detection of peaks, followed by secondary processing to extract the required information on the qualitative and quantitative composition of the substance being analyzed.

## 1. Preliminary Data Analysis

The first stage in processing mass spectra includes detection of peaks on the background of noise, along with surges, base line drift, and effects due to incomplete peak resolution, with conversion of the continuous spectrum into a discrete spectrum. This is carried out using a search method based on matched filtering or comparison of derivatives at three points of a sliding data window [4]. Both approaches can be used independently or sequentially (the first after the second) to supplement each other.

Mass spectra can be represented as an additive mix of $K$ peaks $X_k(t)$, $k = 1, ..., K$, for example of Gaussian form, noise $n(t)$, and the baseline $f(t)$:

$$y(t) = \sum_{k=1}^{K} x_k(t) + n(t) + f(t), \quad (1)$$

where

$$x_k(t) = A_k e^{-\frac{t-t_k}{2w_k^2}};$$

$t$ is continuous or discrete time, $t = i\delta t$ ($i = 1, 2, ..., N$), $N$ is the total number of mass spectral samples, $\delta t$ is the digitization interval, $A_k$, $t_k$, and $w_k$ are the intensity, position, and mean square width of the $k$th peak, respectively, $n(t)$ is noise, and $f(t)$ is baseline.

The *first method* uses a sliding convolution of the initial signal $y(t)$ with a function fitting peak shape:

$$s(t) = \int_{t-T}^{t+T} y(t_1)\, g(t - t_1) dt_1, \qquad (2)$$

where $t$ is the independent variable (time), $g(t)$ is peak shape, and $T$ is the size of half the sliding data window.

Peak shape in the present studies was defined by two normalized functions:

- the first is based on a Gaussian:

$$g_k(t) = e^{-\frac{t}{2w^2}},$$

where $w$ is the peak halfwidth;

- the second is based on the Hermite function:

$$\psi(x) = \frac{e^{-\frac{t^2}{2}}}{\sqrt{2^n n! \sqrt{\pi}}} H_n(t),$$

where

$$H_n(t) = (-1)^n e^{t^2} \frac{d^n e^{-t^2}}{dt^n}$$

is the Hermite polynomial of order $n$ at $n = 2$.

Filtration increases the signal-to-noise ratio by factors of 6-7 for Gaussian functions and 7-10 for Hermite functions. However, the Hermite function has the advantage over Gaussian functions in increasing the resolution of poorly resolved peaks. The convolution procedure provides for reliable detection of spectral peaks and determination of their positions.

The signal $s(t)$ is compared with threshold $h$ used for peak detection and positioning. If the value of $s(t)$ is greater than $h$, this point $t$ is regarded as a peak. These algorithms have been described in more detail in [5, 6].

The *second method* uses a traditional approach to detecting peaks in terms of transection of the first derivative of the null line. Detection reliability for the beginning, end, peak, and interpeak valley can be increased by comparing derivatives at three points of the sliding data window. Different methods for evaluating peak parameters using derivatives have been described in [4, 7-9].

## 2. Secondary Data Processing

The next stage in processing was classification of the processed mass spectra. Discriminant and cluster analysis led to classification into groups. The mass spectra from groups of healthy and sick patients were notably different, particularly in terms of the latter group having additional components with masses 55, 59, 64, and 71 Da. This feature is evidence of changes in the body induced by particular diseases. It can be suggested that the ratios of the intensities of mass spectral components contain information on the type of pathology, which, after selection of the appropriate statistics, provides for effective diagnosis of pathology and, furthermore, classification into groups of diseases.

The "healthy patients / intensity of mass spectral components" dataset has the "object–sign" matrix $XS = |x_{ij}|$, where $i = 1, ..., N$ is the number of objects (patients); $j = 1, ..., M$ is the number of signs – mass numbers; $x_{ij}$ is the intensity of the $j$th spectral component of the $i$th healthy patient.

The dataset of patients forms an $I_0 \times J$ matrix in which the rows contain the intensities of the mass components of exhaled air from one patient. This matrix is designated $XI = |x_{ij}^0|$, where $i = 1, ..., I_0$ is the number of objects (patients); $j = 1, ..., J$ is the number of signs, i.e., mass numbers; $x_{ij}^0$ is the intensity of the $j$th spectral component of the $i$th sick patient. The matrices "patient / mass spectra of healthy and sick people" are used to form a single matrix $X$ of size $(I_0 + I) \cdot J$: $X = (|XS, XI|)$.

Using these data, we form a space of principal components (PC) and calculate the matrix of scores $T$.

We process matrix $T$ with algorithms for discriminant and hierarchical agglomerative cluster analyses.

**2.1. Discriminant analysis.** Discriminant analysis is an approach to multidimensional statistical analysis that includes methods for classifying multidimensional observations on the principles of maximal similarity between the observations being analyzed and observations belonging to defined classes on the basis of training results. Determination of coefficients of discriminant functions uses QR decomposition or the algorithms described in [10]. Mass spectra from healthy people were selected at the preliminary processing stage for inclusion in the training set.

The need to carry out this step is explained by the fact that even the mass spectra of essentially healthy people can contain spectral lines with amplitudes character-

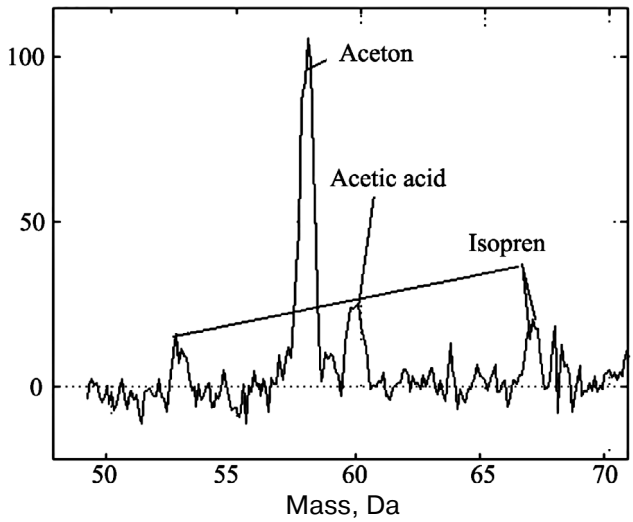


**Fig. 1.** Standard mass spectrum (the vertical scale shows peak intensity, relative units).

istic of the mass spectra of sick people. This step was run by identifying mass spectra containing the smallest deviations from the standard mass spectra, i.e., averaged mass spectra from healthy people. Figure 1 shows the standard mass spectrum.

Training results consisted of a set of coefficients for the discriminant function, which were calculated using one of the algorithms discussed in [10]. Variables in multidimensional statistics algorithms did not use the initial peak amplitudes for defined masses but the principal components obtained by transformation of the initial data by the principal component analysis (PCA).

The calculated discriminant functions had boundaries separating the initial mass spectra of exhaled gases into groups: healthy people and people with pathology. Figure 2 shows results obtained by discriminant analysis using the linear (Fig. 2a) and quadratic (Fig. 2b) methods, respectively.

Processing of the mass spectra of exhaled gases by discriminant analysis showed the probability of type I errors to be tending to zero.

**2.2. Cluster analysis.** Data processing consisted of two stages: *training* and *classification*. At the *training* stage, the spectra of healthy patients formed a space of PC, with calculation of eigenvectors $P$ and eigenvalues $\lambda = [\lambda_1, \lambda_2, ..., \lambda_J]$ of covariation matrix $K$ and computation of the projections of the learning samples on the axis of the principal components matrix: $T = \overset{0}{XS} \times P$, where $\overset{0}{XS}$ is the centered mass spectrum for healthy patients.

At the *diagnosis* stage, the mass spectra of the study patients are analyzed.

1. The projections of the mass spectrum of the patient on the principal components axis are computed: $T_c = \overset{0}{XI} * P$, where $\overset{0}{XI}$ is the centered mass spectrum of the patient. Centering of vector $XI$ is performed using the mean values for columns of matrix $X$.

Hierarchical algorithmic clustering is applied to matrix $T_c$, which is required for separating objects into clusters. This yields several clusters of data characterizing

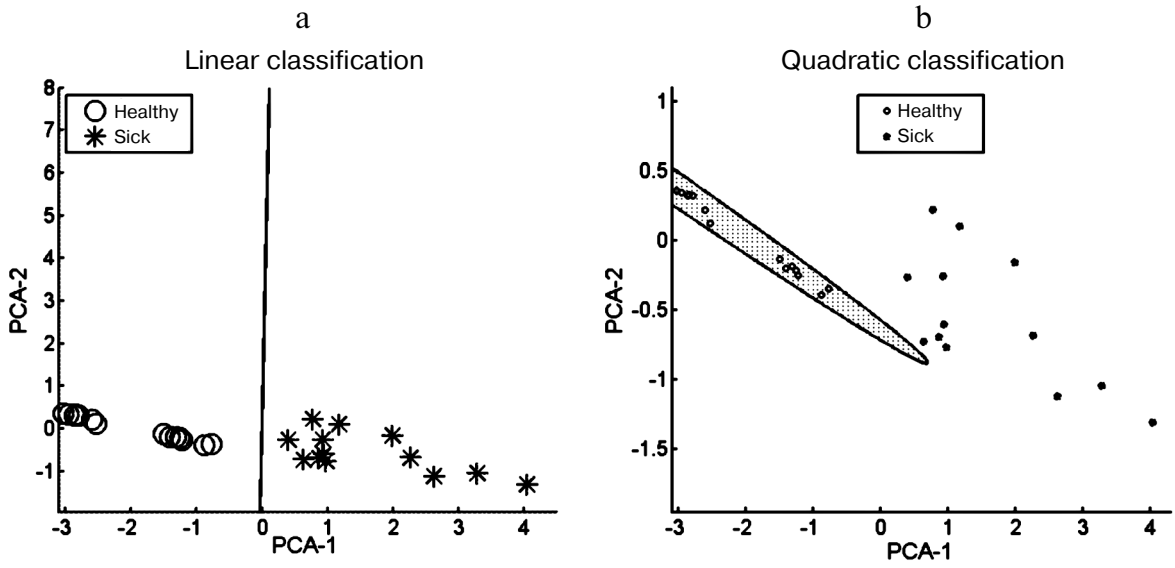


**Fig. 2.** Results of discriminant analysis: the horizontal axis shows variables for the first principal component (PCA-1); the vertical axis shows variables for the second component; a) results of linear discriminant analysis; b) results of quadratic discriminant analysis.

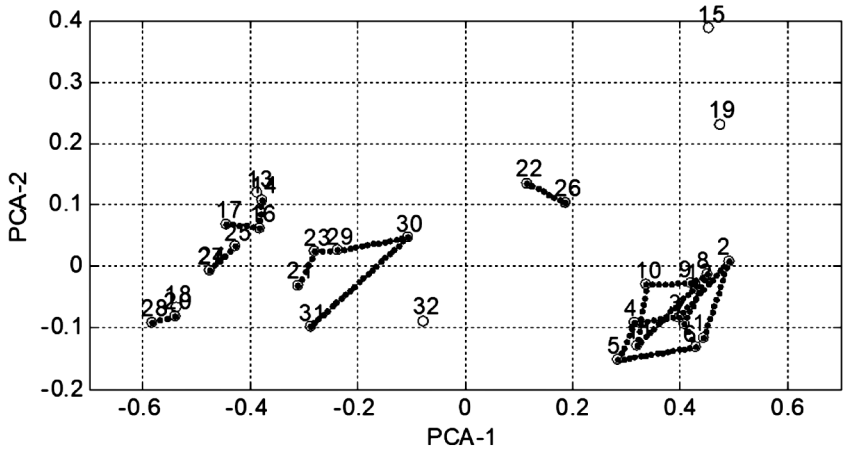


**Fig. 3.** Classification of mass spectra in terms of the state of health by cluster analysis. Numbers on plots show numbers of mass spectra of exhaled gas from healthy and sick patients.

the specific features of patients' diseases. Figure 3 shows separation of healthy (1-12) and sick (20-32) patients into clusters. Sick patients were divided into eight groups: 13, 14, 16, 17; 15; 18, 20, 28; 19; 21, 23, 29, 30, 31; 22, 26; 24, 25, 27; 32. PCA-1 and PCA-02 are the first and second principal components axes.

During subsequent processing, the probability of correct classification was calculated. For a patient belonging to the class of healthy, this probability is maximum when the Euclidean distance between the center of the "cloud" (the centroid) and the spectral point of this patient in the PC space is close to zero.

The algorithm has been described in more detail in [11, 12].

## Conclusions

The algorithms considered here for preliminary processing and classification based on discriminant and cluster analysis provide for automatic decision-making regarding differences in mass spectra without visual information analysis presented graphically. This automatic decision-making may be useful for large-scale rapid analysis. When there are minor deviations in the calculated values of discriminant functions, repeat measurements should be obtained with additional information analysis. Furthermore, these methods should be used in parallel with other methods for analysis of pathology.

These algorithms are simple to run and provide for automated decisions regarding the assignment of a signal to a class.

The algorithms have practical value for specialized small, low-cost mass spectrometers, where use of automated classification into health groups using mathematical methods is particularly appropriate.

This work was supported by the State Contract No. 075-00780-19-00.

## REFERENCES

1. Kuz'min, A. G., A Quadrupole Mass Spectrometer [in Russian], RF Utility Model Patent No. 94763, May 27, 2010.
2. Kuz'min, A. G. and Titov, Yu. A., "Small mass spectrometers for dynamic studies of the composition of exhaled air," in: Proc. I International Scientific and Applied Conference "High-Tech Basic and Applied Research in Physiology and Medicine", November 18-19, 2010, Part 3, St. Petersburg, pp. 266-270.
3. Kuz'min, A. G., Tkachenko, E. I., Oreshko, L. S., and Titov, Yu. A., "Perspectives of a method for mass spectrometric aroma diagnosis using the composition of exhaled air," in: Proc. X Eurasian Scientific Conference "Prenosology 2014," December 18-19, 2010, St. Petersburg, pp. 229-231.
4. Novikov, L. V. and Kurkina, V. V., "A method for assessing the parameters of spectral peaks," Nauchn. Priborostr., **27**, No. 3, 99-106 (2017).
5. Vivó-Truyols, G., Torres-Lapasió, J. R., van Nederkassel, A. M., van der Heyden, Y. V., and Massart, D. L., "Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals. Part I: Peak detection," J. Chromatogr. A, **1096**, 133-145 (2005).
6. Manoilov, V. V., Kuz'min, A. G., Zarutskii, I. V., Titov, Yu. A., and Samsonova, N. S., "Methods for processing and investigating the potential of the mass spectra of exhaled gases," Nauchn. Priborostr., **29**, No. 1, 106-109 (2019).
7. Fredriksson, M. J., Petersson, P., Axelsson, B. O., and Bylund, D., "An automatic peak finding method for LC-MS data using

Gaussian second derivative filtering," J. Sep. Sci., **32**, 3906-3918 (2009).
8. Gregoire, J. M., Dale, D., and Bruce van Dover, R., "A wavelet transform algorithm for peak detection and application to powder x-ray diffraction data," Rev. Sci. Instrum., **82**, 015105-015112 (2011).
9. Du, P., Kibbe, W. A., and Lin, S. M., "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching," Bioinformatics, **22**, 2059-2065 (2006).
10. Manoilov, V. V., Titov, Yu. A., Kuz'min, A. G., and Zarutskii, I. V., "Algorithms for discriminant analysis for classification of the mass spectra of exhaled gases," Nauchn. Priborostr., **27**, No. 3, 33-57 (2017).
11. Novikov, L. V. and Kurkina, V. V., "Multidimensional processing of data from mass spectrometric analysis of the composition of exhaled air," in: Proc. XXIX International Scientific Conference "Mathematical Methods in Techniques and Technologies" (MMTT-29) in 12 Volumes [in Russian], A. A. Bol'shakov (ed.), Saratov Sate Technical University, Saratov; St. Petersburg State Institute of Technology (Technical University), St. Petersburg Polytechnical University, St. Petersburg Institute of Informatics and Automation, St. Petersburg; Samara State Technical University, Samara; Vol. 9, pp. 32-38.
12. Berikov, V. B. and Lbov, G. S., Current Trends in Cluster Analysis [in Russian], S. L. Sobolev Institute of Mathematics, Siberian Branch, Russian Academy of Sciences (2008); https://reslib.org/books/642840.