Available online at www.sciencedirect.com

ScienceDirect

Sensors and Actuators B 120 (2007) 467–472

SENSORS AND ACTUATORS B CHEMICAL

www.elsevier.com/locate/snb

A feature extraction algorithm for multi-peak signals in electronic noses

R. Haddad, L. Carmel[1], D. Harel*

Department of Computer Science and Applied Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel

Received 14 June 2005; received in revised form 27 December 2005; accepted 27 February 2006
Available online 18 April 2006

Abstract

The Lorentzian model is a powerful feature extraction technique for electronic noses. In a previous work, it was applied to single-peak transient signals and was shown to achieve lower classification error rate than other feature extraction techniques. Here, we generalize the Lorentzian model by showing how to apply it to transient signals that are comprised of more than a single peak. The model is based on a fast and robust fitting of the measured signals to a physically meaningful analytic curve. We show that this model fits equally well to sensors of different technologies and embeddings, suggesting its applicability to a diverse repertoire of sensors and analytic devices.

Keywords: Feature extraction; Electronic nose; Signal processing; Multiple peaks

1. Introduction

Retrieving information from large datasets usually involves a feature extraction stage, aimed at reducing data dimensionality. A good feature extraction technique is measured by how well the condensed representation preserves the information content of the original data. Electronic noses (or, in short, eNoses) are analytic devices that play a constantly growing role as general purpose detectors of vapor chemicals [5]. The main component of an eNose is an array of non-specific sensors, i.e., sensors that interact with a broad range of chemicals with varying strengths. Correspondingly, an analyte stimulates many of the sensors in the array and elicits a characteristic response pattern.

The sensors inside an eNose are made of diverse technologies. Depending on the type of sensor, a certain physical property is changed as a result of an exposure to gaseous analytes. During the measurement process, a signal is obtained by constantly recording the value of this physical property. A typical eNose signal is comprised of a few hundred measured values per sensor, thus giving rise to a rather large dataset. A preceding stage of feature extraction is therefore almost mandatory. The most commonly used methods (see examples in Fig. 1) capture only a portion of the information contained in the signals. Even though these methods are satisfactory for some applications, it is generally accepted that performance can be enhanced by the use of more optimal methods.

One such method, the Lorentzian model, was suggested in [1]. Additional proposals can be found, e.g., in [4]. The Lorentzian model is based on fitting the measured signal to an analytic curve, developed using simple assumptions regarding the measurement system and the interaction between an analyte and the sensors. The resulting feature extraction technique uses four parameters to characterize each signal, and was shown to significantly outperform other techniques [1]. A demonstration of the fit between the measured signal and the analytic one is shown in Fig.2.

The Lorentzian model assumes single-peak transient signals (see Fig. 2), and is consequently appropriate for ordinary signals obtained by transient measurement. In practice, however, measured response signals occasionally exhibit abnormal signal shapes, thus posing difficulties for the Lorentzian technique. We may classify the abnormal signals into two categories:

(1) *Corrupted signals*: Occur when the sensor (or its supporting electronics) fails at a certain time during the measurement and then recovers and continues measuring; see Fig. 3a, b and d. Corrupted signals can be further classified into three different sub-types, see [2].
(2) *Multi-peak signals*: Occur when the signal exhibits more then one significant peak; see Fig. 3c and d.

* Corresponding author. Tel.: +972 8 9344 050; fax: +972 8 9344 122.
[1] Present address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Building 38A, Bethesda, MD 20894, USA
E-mail addresses: refael.haddad@weizmann.ac.il (R. Haddad), carmel@ncbi.nlm.nih.gov (L. Carmel), dharel@weizmann.ac.il (D. Harel).

0167-0987/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.snb.2006.02.048

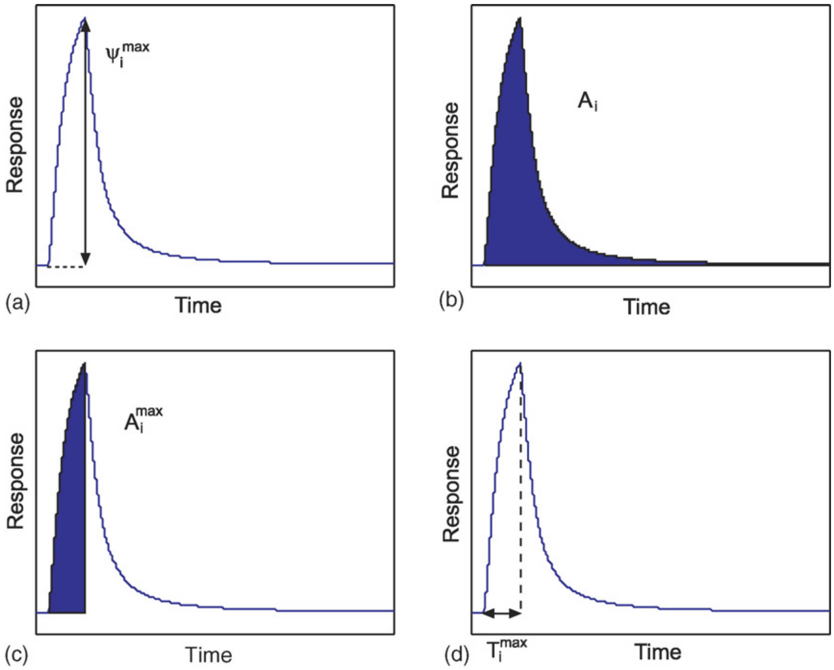

Fig. 1. Definition of the four most popular features in transient signals. (a) The difference between the peak and the baseline, $\psi_i^{\max}$. (b) The area under the curve, A_i. (c) The area under the curve left of the peak, $A_i^{\max}$. (d) The time from the beginning of the signal to the peak $T_i^{\max}$.

In its original form, the Lorentzian feature extraction technique cannot be applied to abnormal signals. Nevertheless, it was later shown that damaged parts of corrupted signals can be restored [2], making such signals appropriate for application of the Lorentzian feature extraction technique. Still, multi-peak signals kept defeating most feature extraction techniques, and were usually left outside of the analysis. In this paper we aim at changing this situation by suggesting a generalization of the Lorentzian feature extraction technique, enabling it to be applied to multi-peak signals as well.

To this end, it is beneficial to interpret a multi-peak signal as if each peak is produced by a different subset of components of the incoming stimulus. We can further assume that each individual peak has the typical Lorentzian shape described in [1], such that the overall signal is just a superposition of Lorentzian signals. We show that such a model yields excellent fits to measured signals, and gives rise to an informative and powerful feature extraction technique.

Our work renders the Lorentzian feature extraction technique applicable to any kind of transient signal obtained in the laboratory. In contrast, most feature extraction techniques that we are aware of fail in at least one of the abnormal signal classes. This fact further broadens the scope and applicability of our method.

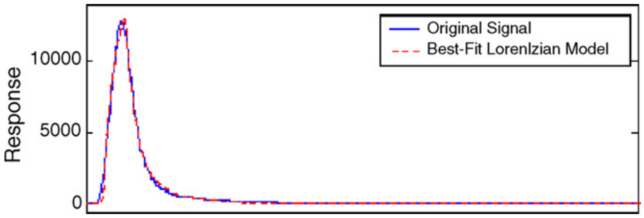

Fig. 2. A typical signal (*cis*-3-hexenyl acetate) measured with a QMB sensor and the result of using the Lorentzian model.

2. Experimental

We have tested our algorithm using data collected by the MOSESII eNose [7] with two sensor modules: an eight-sensor quartz-microbalance (QMB) module, and an eight-sensor metal-oxide (MOX) module. The samples were put in 20-ml vials in an HP7694 headspace sampler, which heated them to 40 °C and injected the headspace content into MOSESII. There the analyte was first introduced into the QMB chamber, whence it followed to the 300 °C heated MOX chamber. The injection lasts 30 s, and is followed by a 15 min purging stage using synthetic air.

The dataset comprised 70 volatile odorous pure chemicals, intentionally chosen from many different chemical families, so that they would represent a broad range of possible stimuli. Each chemical was measured in batches, with a single batch containing at least seven successive measurements. In total, we performed 675 measurements. Of the 70 chemicals measured, 54 had their sensors properly responding; whereas 16 (∼20%) had at least one signal with more than a single peak. Interestingly, the multi-peak phenomenon is twice as abundant in QMB signals than in MOX.

3. The generalized lorentzian model

As mentioned earlier, we adopt the interpretation that each peak in the signal ensue from a different subset of mixture components. Theses subsets are probably characterized by significantly different volatilities, causing them to exhibit a kind of chromatographic effect in their path through the eNose's

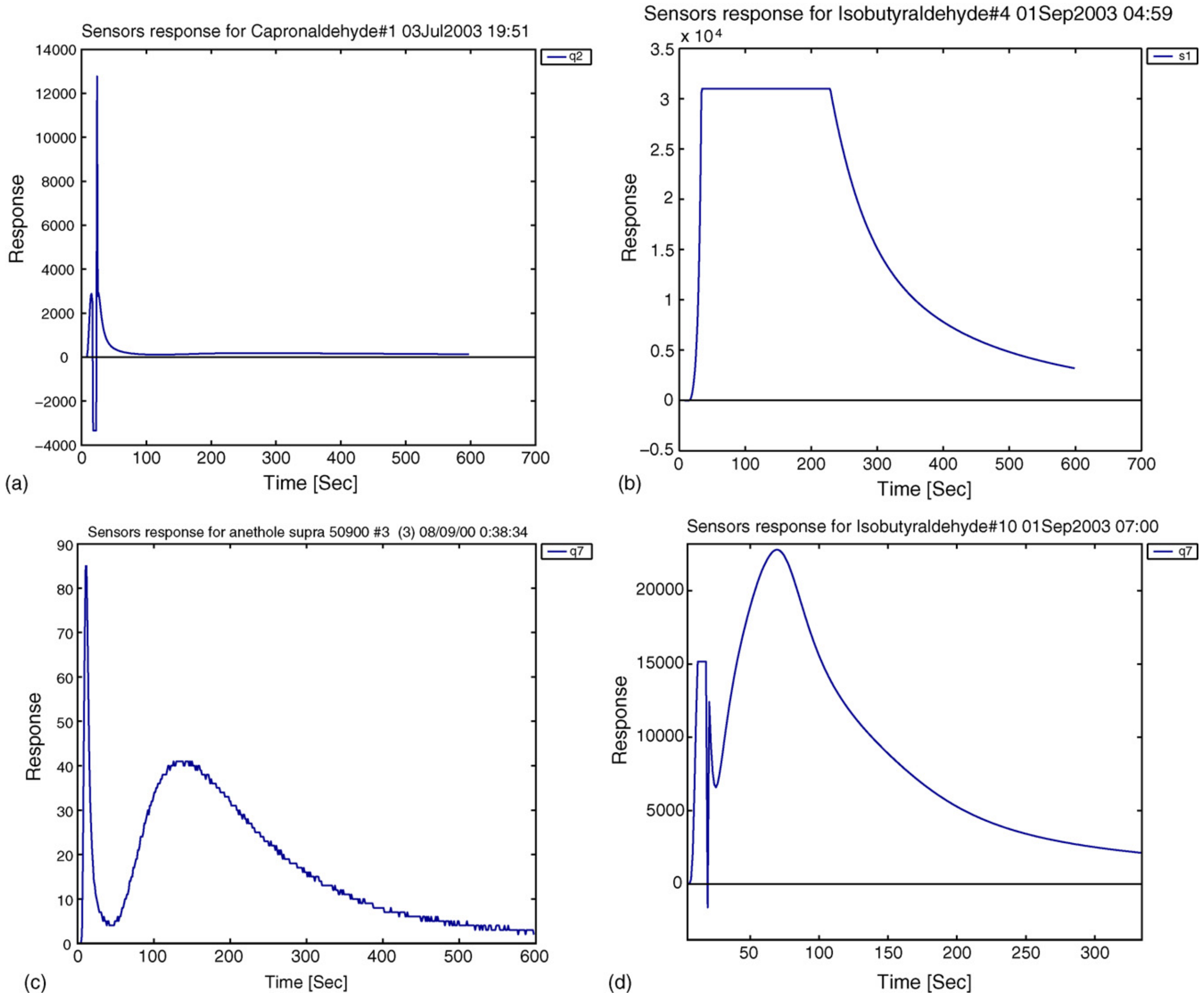

Fig. 3. Abnormal signals. (a) A failure in a measurement. Note that the corruption is only temporary, and afterwards the signal resumes its typical behavior. (b) A failure in a measurement, realized as a Plateau indicating electronic saturation of the measurement system. Again, after a while, the signal resumes its typical behavior. (c) A double-peak signal. (d) A measurement exhibiting all kinds of phenomena—failure, saturation and double-peak.

pipeline. In fact, each such subset of the mixture components need not be a pure chemical, and may be a mixture in itself.

For simplicity, we limit the following discussion to the case of two peaks, but the results can be readily generalized to more peaks. Interestingly, in our case, 99.9% of the multi-peak signals obtained were actually double peaks, so that only in a handful of cases did we have to deal with more than two peaks.

In [1] a simple physical description of the measurement system was used to derive the analytic expression for the shape of the response signal, explicitly given by

$$L(t;\theta)=\begin{cases}0, & t<t^0,\\ \beta\tau\,\tan^{-1}\left(\frac{t-t^0}{\tau}\right), & t^0\le t\le t^0+T,\\ \beta\tau\left[\tan^{-1}\left(\frac{t-t^0}{\tau}\right)-\tan^{-1}\left(\frac{t-t^0-T}{\tau}\right)\right], & t>t^0+T.\end{cases} \tag{1}$$

Here, t^0 is the time when the signal starts to rise, T the time interval between the signal rise and its peak, τ a characteristic of the signal's decay time, β relates to its amplitude, and $\theta=\{\beta,\tau,t^0,T\}$ represents the entire set of parameters.

A linear decomposition of two Lorentzian signals would then be

$$L(t;\theta)=L_1(t;\theta_1)+L_2(t;\theta_2), \tag{2}$$

where L_1 and L_2 are the Lorentzian signals describing each of the peaks, and $\theta=\theta_1\cup\theta_2$, $\theta_i=\{\beta_i,\tau_i,t_i^0,T_i\}$, $i=1,2$, is the set of all parameters. It is hereinafter assumed that L_1 is the earlier (left hand-side) peak, while L_2 is the later (right hand-side) peak. Our assumption states that $L_1(t;\theta_1)$ and $L_2(t;\theta_2)$

depict the Lorentzian signals that would have been obtained had we measured the low-volatiles and the high-volatiles separately.

3.1. Implementation

The parameter set θ is found by fitting the analytic model (2) to the measured signal. Since this function is not everywhere differentiable, we could not use gradient based methods for the curve fitting, and preferred the Matlab function *fminsearch*, which uses the simplex method [6]. It is the custom in curve fitting to minimize the sum of squared differences between the measured signal and the analytic one. However, since a typical transient eNose signal has a relatively long decaying part (in most cases more than half of the signal duration, see example in Fig. 2), we used a weighted cost function for the minimization, giving the points before the decay part twice the weight.

To this end, we have divided the signal s into two parts s_1 and s_2, where s_1 represent the values of the signal from the start until the decay of the second signal and s_2 represent the rest of the signal. We then compute the best-fitting single-peak Lorentzian model, and, based on this function we define l_1 and l_2, where l_1 is the calculated Lorentzian function for the first signal part and l_2 for the remaining part. The weighted cost function formula was therefore,

$$w = 2(s_1 - l_1)^2 + (s_2 - l_2)^2.$$

This modification significantly improves the convergence rate of the curve fitting algorithm.

The speed of convergence and accuracy of the solution are susceptible to the initial values that we assign to the different parameters. As all the parameters are physically meaningful, we are able to supply a rather accurate initial guess, based on the following procedure:

- Estimating T_i and t_i^0 for $i = 1, 2$. Due to the superimposition of the two signals, finding T_i and t_i^0 is somewhat subtle. In the following, t and h stand for time and signal height, respectively. The first step is to find the rising time of the entire signal, t_1, which is assumed to be the rising point of $L_1(t; \theta_1)$. Then, we find the points where the first and second signals obtain their maximum,

  $$P_i^{\max} = (t_i^{\max}, h_i^{\max}), \quad i = 1, 2.$$

  Clearly, these maximum points are not identical to the maximum points of L_1 and L_2. These, as well as the rising point of the second signal, are inferred from the a simple linear extrapolation procedure shown in Fig. 4. Using this linear extrapolation we find that the value of L_1 at $t_2^{\max}$ and the value of L_2 at $t_1^{\max}$. We mark these points as E_i

  $$E_i = (t_i^{\max}, ns_i), \quad i = 1, 2,$$

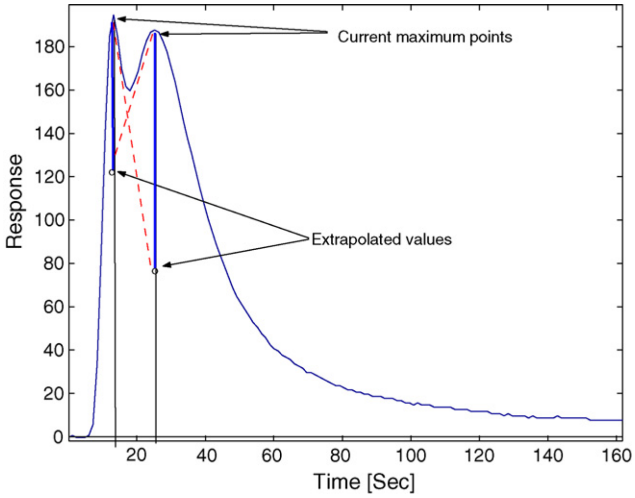

Fig. 4. Extrapolating the original maximum values. The lengths of the blue lines are the new extrapolated maximum values.

  where ns_i is the extrapolated height of L_i for $i = 1, 2$. The new maximum values are calculated by subtracting the extrapolated points E_i from the current maximum points P_i (see Fig. 4).

- Estimating τ_i for $i = 1, 2$. We have estimated τ_1 and τ_2 following the same strategy as in [1], namely averaging over approximated results from the entire dataset. The results for each of the 16 sensors are given in the Table 1. It should be noted that the estimation of τ_i is data-specific, and the values in Table 1 will have to be recomputed for any dataset with is essentially different from ours.

  The decay time of the second signal is always significantly larger (signifying a slower decay) than that of the first signal. This can be clearly seen both from the table and from the example given in Fig. 3c and d. This difference in decay rates can be explained by the fact that the second signal corresponds to a heavier stimulus, resulting in slower rise time and decay time. Note that τ_1 values are not the same as the ones we used in [1]. This is because the decay of the lighter signal is slowed due to the presence of the second signal.

- Estimating β_i for $i = 1, 2$. For β_i we use the same formula as in [1],

  $$\beta_i = \frac{s_i^{\max}}{\tau_i \tan^{-1}(T_i/\tau_i)}, \quad i = 1, 2.$$

4. Results

A sense of how well is the fit that we get can be found in Fig. 5, which shows two examples of generalized Lorentzian fits to measured signals.

Table 1
Initial values used for τ_1 and τ_2 that are used as inputs for the curve fitting process

	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
τ_1	10	10	5	5	10	7	7	5	40	20	50	40	30	40	30	50
τ_2	70	100	40	80	120	166	90	80	93	250	250	250	200	150	200	300

Table 2
Averages and medians of the R^2-test, applied to our curve fitting process, for theo analytic models

Sensor	Lorentzian		Exponential	
	Average	Median	Average	Median
Q_1	0.9943	0.9954	0.9901	0.9915
Q_2	0.9927	0.9929	0.9898	0.9910
Q_3	0.9937	0.9947	0.9720	0.9876
Q_4	0.9942	0.9947	0.9892	0.9901
Q_5	0.9934	0.9943	0.9908	0.9928
Q_6	0.9915	0.9908	0.9881	0.9889
Q_7	0.9945	0.9959	0.9751	0.9925
Q_8	0.9964	0.9974	0.9915	0.9944
S_1	0.9893	0.9929	0.9664	0.9894
S_2	0.9897	0.9904	0.9460	0.9870
S_3	0.9893	0.9931	0.9257	0.9906
S_4	0.9889	0.9935	0.8928	0.9951
S_5	0.9933	0.9964	0.9693	0.9935
S_6	0.9894	0.9919	0.9310	0.9824
S_7	0.9890	0.9909	0.9588	0.9827
S_8	0.9908	0.9914	0.9715	0.9871

Q_1–Q_8 are the eight QMB sensors, and S_1–S_8 are the eight MOX sensors. For all sensors, whether QMB or MOX, the Lorentzian model gives superior R^2-values, although both models are quite good.

To quantify how well a model fits the data, we used the well known R^2-test [3] for goodness-of-fit. This test is bounded from above by 1, and the closer it gets to 1, the better the fit in the least squares sense. The advantage of the R^2-test is that it measures goodness-of-fit on a normalized scale, thus enabling comparison between differently scaled signals. We tested our model against all 300 × 8 QMB signals, and 300 × 8 MOX signals, and calculated the average and the median of R^2. This time we did not use a weighted cost function as we want to compare how well the resulting signals fit the original ones. The results are shown in Table 2.

To evaluate the quality of the results, we compared them to those that are obtained by using the exponential model instead of the Lorentzian model. The exponential model was developed in [1] as an alternative to the Lorentzian model; it is given by

$$E(t;\theta)=\begin{cases}0, & t<t^0,\\ \beta\tau(1-\mathrm{e}^{-(t-t^0/\tau)}), & t^0\le t\le t^0+T,\\ \beta\tau(\mathrm{e}^{T/\tau}-1)\,\mathrm{e}^{-(t-t^0/\tau)}, & t>t^0+T.\end{cases} \tag{3}$$

Again, the entire signal is assumed to be a superposition of two exponential signals. In a way, the exponential model is more natural in that it assumed the familiar exponential decay, but in [1], the Lorentzian model was shown to yield better classification. Here, too, we show that the generalized Lorentzian model is preferred to the generalized exponential model; see Table 2.

As eNoses are mostly used for the purpose of classification, it is a good practice to test how well our feature extraction technique allows to discriminate between different odorous mixtures. In Table 3 we compared the success rate of three classification methods using two versions of feature extraction techniques: the popular signal height (taking the height of the highest peak), and our generalized Lorentzian model. As can be seen from the this table, the generalized Lorentzian model gives higher classification rates.

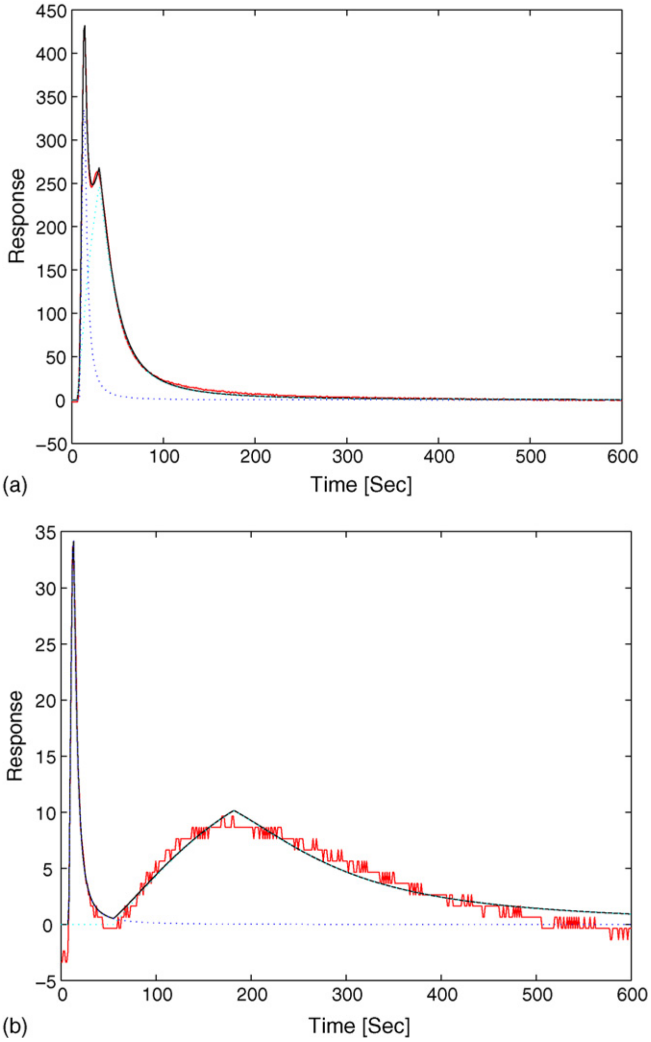

Fig. 5. Two examples of the Lorentzian model fitted to a measured double peak signal. The red line depicts the original signal, while the black depicts the best fitting generalized Lorentzian model. The two dashed lines show the two individual Lorentzian signals that are superimposed to reconstruct the measured double peak singal.

Table 3
The success rate of three different classification methods using the two feature extraction methods discussed in the text

	Signal height (%)	Generalized Lorentzian (%)
Bayes	63.94	72.72
KNN	68.68	70.60
Perceptron	66.05	84.53

As can be seen, the generalized Lorentzian model gives higher classification rates. The analysis was carried out on a set of 390 odor signal samples, measured using the MOSESII electronic nose. The data was classified into two groups according to some specific odor property.

5. Discussion

We have reason to believe that a multi-peak signal occurs when the input mixture can be divided into disjoint sub-mixtures,

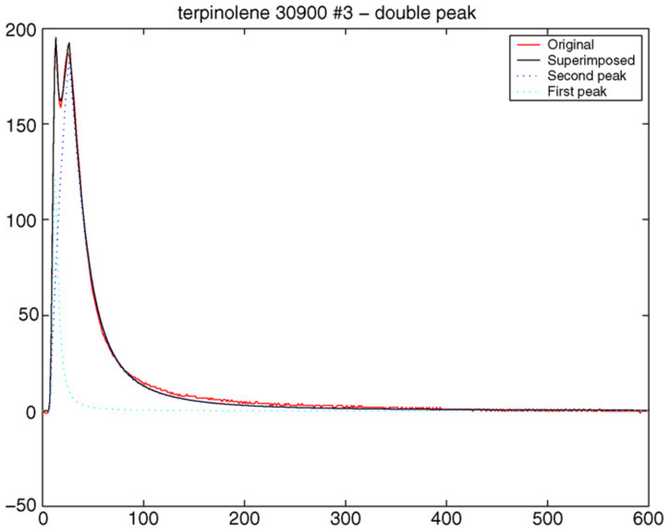

Fig. 6. Double peaks with two fitting Lorentzians. The dashed blue curve is the first signal and the dashed green one is the second, the red curve is the superposition of the two curves. Note that it is hard to tell if this is a true double peak signal or a signal with a failure section.

having significantly different volatilities. Previously, the common practice was to discard such signals prior to the application of any data analysis technique, thus potentially loosing valuable information. The present work, combined with the results in [2] on remedying corrupted signals, enable the utilization of non-standard signals in the data analysis.

After a multi-peak signal has been decomposed into its single-peak constituents, it is left to the user to decide which of them to use for the analysis. In many cases, one of the peaks is caused by some contamination of the sample, in which case the stronger peak should be used for subsequent analysis. In other cases, a very rapid first peak is caused by large values of humidity in the sample and then comes the following signal, which carries the important information. Sometimes, we suggest comparing the results with other datasets as a strategy to decide on the nature of the two peaks. In the absence of any external knowledge, we may suggest to use both peaks in the analysis and to examine which of them gives a result that is more consistent with the other results.

Whenever one of the peaks reflects the impact of an undesirable element (like contamination or humidity), its removal will result in a cleaner signal. This might serve as a remedy to the well documented sensitivity to humidity in eNoses.

Identifying multiple peaks is a technical question with broad practical implications. Sometimes, as is demonstrated in Fig. 6, it is hard to decide if the signal under inspection is corrupted or is simply a multiple-peak signal. This situation is quite rare in our dataset, as most of the failures are easily identified (see, e.g., Fig. 3). In these few cases where we have doubts, we can decide on the signal classification using a simple test involving the signal height in the vicinity of the two peaks. We rely on the observation, at least true to the specific dataset used, that each sensor has a typical range of response, for example, the first QMB sensor (Q_1) usually has its peak in the range of 1900–2200. A failure occurs when the sensor is driven above its “normal” operating range. Therefore, if the double peak occurs for very high readings of the sensor, we assume that it is a failure, while if it occurs for low readings, we regard it as a double-peak. For example, in Fig. 6, the abnormal signal shape is associated with a double peak and not with a failure.

References

[1] L. Carmel, S. Levy, D. Lancet, D. Harel, A feature extraction method for chemical sensors in electronic noses, Sens. Actuators B: Chem. 93 (2003) 67–76.

[2] L. Carmel, Electronic nose signal restoration: Beyond the dynamic range limit, Sens. Actuators B: Chem. 106 (2005) 95–100.

[3] W.R. Dillon, M. Goldstein, Multivariate Analysis Methods and Applications, John Wiley and Sons, New York, USA, 1984.

[4] C. Distante, M. Leo, P. Siciliano, K.C. Persaud, On the study of feature extraction methods for an electronic nose, Sens. Actuators B: Chem. 87 (2002) 274–288.

[5] J.W. Gardner, P.N. Bartlett, Electronic Noses, Principles and Applications, Oxford University Press, New York, USA, 1999.

[6] The Math Works Inc., Optimization Toolbox for use with Matlab, User Guide Version 2, 4th Printing (Release 12), 2000.

[7] J. Mitrovics, H. Ulmer, U. Weimar, W. Gopel, Modular sensor systems for gas sensing and odor monitoring: the MOSES concept, Acc. Chem. Res. 31 (1998) 307–315.

Biographies

Rafi Haddad received his BSc in Mathematics and Computer Science from Tel-Aviv University in 1995, and his MSc degree in Computer Science and Applied Mathematics, specializing in bioinformatics, from the Weizmann Institute of Science in 2005. He recently started his PhD studies in the same department. His research deals with materializing odor digitization, transmission and reproduction, and it involves many kind of mathematics (e.g., multivariate data analysis and statistical pattern recognition), biology (e.g., the sense of smell and receptor repertoires), and chemistry (e.g., electronic noses and chemical sensors).

Liran Carmel received his PhD degree in 2003, in the Department of Computer Science and Applied Mathematics at the Weizmann Institute of Science, Israel. In his research, he investigated the feasibility of odor communication, and developed many algorithms for this purpose, including a classification algorithm, a feature extraction technique, algorithms for dimensionality reduction, and algorithms for multivariate data visualization. Currently, he is pursuing his postdoctoral research at the National Institutes of Health, USA, where he deals with different aspects of molecular evolution.

David Harel has been at the Weizmann Institute of Science since 1980. He was Department Head from 1989 to 1995, and was Dean of the Faculty of Mathematics and Computer Science between 1998 and 2004. He is also co-founder of I-Logix, Inc. He received his PhD from MIT in 1978, and has spent time at IBM Yorktown Heights, and at Carnegie-Mellon and Cornell Universities. In the past he worked mainly in theoretical computer science, and now he works in software and systems engineering, modeling biological systems, and the synthesis and communication of smell. He is the inventor of statecharts and co-inventor of live sequence charts, and co-designed Statemate, Rhapsody and the Play-Engine. He received the ACM Outstanding Educator Award in 1992 and the Israel Prize in 2004.